



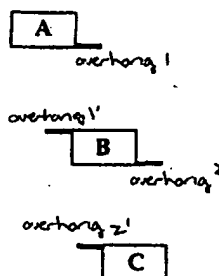
INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁷ : C12N 15/11, 15/62, C12Q 1/68, C07H 21/00		A2	(11) International Publication Number: WO 00/40715
			(43) International Publication Date: 13 July 2000 (13.07.00)
(21) International Application Number: PCT/US00/00189		(74) Agent: JARRELL, Brenda, Herschbach; Choate, Hall & Stewart, Exchange Place, 53 State Street, Boston, MA 02109 (US).	
(22) International Filing Date: 5 January 2000 (05.01.00)			
(30) Priority Data: 09/225,990 5 January 1999 (05.01.99) US 60/114,909 5 January 1999 (05.01.99) US		(81) Designated States: AU, CA, JP, US, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).	
(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application US Not furnished (CIP) Filed on Not furnished		Published <i>Without international search report and to be republished upon receipt of that report.</i>	
(71) Applicant (for all designated States except US): TRUSTEES OF BOSTON UNIVERSITY [US/US]; 108 Bay State Road, Boston, MA 02215 (US).			
(72) Inventors; and			
(75) Inventors/Applicants (for US only): JARRELL, Kevin, A. [US/US]; 3 Acorn Lane, Lincoln, MA 01773 (US). COL-JEE, Vincent, W. [US/US]; 119 Pearl Street, Cambridge, MA 02139 (US). DONAHUE, William [US/US]; Apartment 2, 63 Kendall Street, Quincy, MA 02169 (US). MIKHEEVA, Svetlana [US/US]; 1144 Commonwealth Avenue, Apt. 12A, Allston, MA 02134 (US).			

(54) Title: **IMPROVED NUCLEIC ACID CLONING**

(57) Abstract

The present invention provides an improved system for linking nucleic acids to one another. In particular, the present invention provides techniques for producing DNA product molecules that may be easily and directly ligated to recipient molecules. The product molecules need not be cleaved with restriction enzymes in order to undergo such ligation. In preferred embodiments of the invention, the DNA product molecules are produced through iterative DNA synthesis reactions, so that the product molecules are amplified products. The invention further provides methods for directed ligation of product molecules (i. e., for selective ligation of certain molecules within a collection of molecules), and also for methods of exon shuffling, in which multiple different product molecules are produced in a single ligation reaction. Preferred embodiments of the invention involve ligation of product molecules encoding functional protein domains, particularly domains naturally found in conserved gene families. The inventive DNA manipulation system is readily integrated with other nucleic acid manipulation systems, such as ribozyme-mediated systems, and also is susceptible to automation.

**LIBRARY ASSEMBLY USING
RNA/DNA CHIMERIC OLIGOS****COMBINATORIAL POTENTIAL**

$$10 \times 3 = 30$$

$$10^3 = 1000$$

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

IMPROVED NUCLEIC ACID CLONING

5 The present application claims priority to co-pending United States provisional applications U.S.S.N. 09/225,990 and U.S.S.N. 60/114,909, each of which was filed on January 5, 1999 and each of which is incorporated herein by reference in its entirety.

Government Funding

10 Some or all of the work described herein was supported by grant number RO1 GM 52409 from the National Institutes of Health and by grant number MCB9604458 from the National Science Foundation; the United States Government may have certain rights in this invention.

Background

15 The Molecular Biology revolution began with the discovery of enzymes that were capable of cleaving double stranded DNA, so that DNA fragments were produced that could be ligated to one another to generate new, so-called "recombinant" molecules (see, for example, Cohen et al., *Proc. Natl. Acad. Sci. USA* 70:1293, 1973; Cohen et al., *Proc. Natl. Acad. Sci. USA* 70:3274, 1973; see also U.S. Patent Nos. 4,740,470; 4,468,464; 4,237,224). The revolution was
20 extended by the discovery of the polymerase chain reaction (PCR), which allowed rapid amplification of particular DNA segments, producing large amounts of material that could subsequently be cleaved and ligated to other DNA molecules
25 (see, for example, U.S. Patent Nos. 4,683,195; 4,683,202; 5,333,675).

Despite the power of these digestion and amplification techniques, however, there remains substantial room for improvement. Reliance on digesting enzymes, called "restriction enzymes", can render molecular biological experiments quite expensive. Moreover, many of the enzymes are inefficient or are only available in
30 crude preparations that may be contaminated with undesirable entities.

At first, it seemed that PCR amplification might itself avoid many of the difficulties associated with traditional cut-and-paste cloning methods since it was thought that PCR would generate DNA molecules that could be directly ligated to

other molecules, without first being cleaved with a restriction enzyme. However, experience indicates that most PCR products are refractory to direct cloning. One possible explanation for this observation has come from research revealing that many thermophilic DNA polymerases (including *Taq*, the most commonly used enzyme) add terminal 3'-dAMP residues to the products they amplify. Invitrogen (Carlsbad, CA) has recently developed a system for direct cloning of such terminally-dAMP-tagged products (TA Cloning Kit®; see U.S. Patent No. 5,487,993) if the molecule to which they are to be ligated is processed to contain a single unpaired 3'-dTMP residue. While the Invitrogen system has proven to be very useful, it is itself limited in application by being restricted to ligation of products with only a single nucleotide overhang (an A residue), and is further restricted in that the overhang must be present at the 3' end of the DNA molecule to be ligated.

There is a need for the development of improved systems for nucleic acid cloning. Particularly desirable systems would allow DNA ligation with minimal reliance on restriction enzymes, would provide for efficient ligation, and would be generally useful for the ligation of DNAs having a wide variety of chemical structures. Optimal systems would even provide for directional ligation (i.e., ligation in which the DNA molecules to be linked together will only connect to one another in one orientation).

Summary of the Invention

The present invention provides an improved system for linking nucleic acids to one another. In particular, the present invention provides techniques for producing DNA product molecules that may be easily and directly ligated to recipient molecules. The product molecules need not be cleaved with restriction enzymes in order to undergo such ligation. In preferred embodiments of the invention, the DNA product molecules are produced through iterative DNA synthesis reactions, so that the product molecules are amplified products.

The inventive system provides techniques and reagents for generating product molecules with 3' overhangs, 5' overhangs, or no overhangs, and further

provides tools for ligating those product molecules with recipient molecules. Where overhangs are employed, the length and sequence of the overhang may be varied according to the desires of the practitioner. Overhang-containing products may be linked to one another by any available means including, for example, enzymatic ligation or transformation into a host cell. For example, molecules containing at least 12 nt overhangs may be annealed to one another and linked together by transformation into *E. coli* without first being ligated (see, for Example, Rashtchian, et al. *Analytical Biochemistry* 206:91, 1992).

The inventive system further provides methods for directed ligation of product molecules (i.e., for selective ligation of certain molecules within a collection of molecules), and also for methods of exon shuffling, in which multiple different product molecules are produced in a single ligation reaction. Preferred embodiments of the invention involve ligation of product molecules encoding functional protein domains, particularly domains naturally found in conserved gene families. Alternative or additional preferred embodiments of the invention involve multi-component ligation reactions, in which three or more nucleic acid molecules are ligated together. In some embodiments, these multiple molecules are linked in only a single arrangement; in others, multiple arrangements can be achieved.

The inventive DNA manipulation system is readily integrated with other nucleic acid manipulation systems, such as ribozyme-mediated systems, and also is susceptible to automation. Specifically, in one aspect, a double stranded DNA molecule with a single stranded overhang comprised of RNA is provided. Additionally, in another aspect, a library of nucleic acid molecules is provided wherein each member of the library comprises 1) at least one nucleic acid portion that is common to all members of the library; and 2) at least two nucleic acid portions that differ in different members of the library, is also provided by the present invention. In a preferred embodiment, each of the nucleic acid portions in the library comprises protein-coding sequence and each library member encodes a continuous polypeptide. In yet another particularly preferred embodiment, each of the variable nucleic acid portions encodes a functional domain of a protein. This functional domain is preferably one that is naturally found in a gene family selected from the group consisting of the tissue plasminogen activator gene family, the

animal fatty acid synthase gene family, the polyketide synthase gene family, the peptide synthetase gene family, and the terpene synthase gene family.

In yet another aspect of the present invention, a method of generating a hybrid double-stranded DNA molecule is provided. This method comprises the steps of 1) providing a first double-stranded DNA molecule, which double-stranded DNA molecule contains at least one single stranded overhang comprised of RNA; 2) providing a second double-stranded DNA molecule containing at least one single-strand overhang that is complementary to the RNA overhang on the first double-stranded DNA molecule; and 3) ligating the first and second double-stranded DNA molecules to one another so that a hybrid double-stranded DNA molecule is produced. In certain preferred embodiments, the method comprises providing and ligating at least three double-stranded DNA molecules.

A further aspect of the present invention includes a method of generating a hybrid double-stranded DNA molecule, the method comprising 1) generating a first double-stranded DNA molecule by extension of first and second primers, at least one of which includes at least one base that is not copied during the extension reaction so that the extension reaction produces a product molecule containing a first overhang; 2) providing a second double-stranded DNA molecule containing a second overhang complementary to the first overhang; and 3) ligating the first and second double-stranded DNA molecules to one another, so that a hybrid double-stranded DNA molecule is produced. In certain preferred embodiments, the method comprises providing and ligating at least three double-stranded DNA molecules.

In still a further aspect of the present invention, a method of generating a hybrid double-stranded DNA molecule is provided, the method comprising: 1) generating a first double-stranded DNA molecule by extension of first and second primers, at least one of which includes at least one potential point of cleavage; 2) exposing the first double-stranded DNA molecule to conditions that result in cleavage of the cleavable primer at the potential point of cleavage, so that a first overhang is generated on the first DNA molecule; 3) providing a second double-stranded DNA molecule containing a second overhang complementary to the first overhang; and 4) ligating the first and second double-stranded DNA molecules to

one another, so that a hybrid double-stranded DNA molecule is produced. In certain preferred embodiments, the method comprises providing and ligating at least three double-stranded DNA molecules.

5

Description of the Drawing

Figure 1 depicts an inventive process for generating DNA product molecules with 3' overhangs.

Figure 2 depicts a process for producing 5' overhangs by hybridizing a template molecule with one or more primers including at least one ribonucleotide primer.

10

Figure 3 depicts an inventive process for generating DNA product molecules with one or more 5' overhangs.

Figure 4 depicts an alternative inventive process for generating DNA product molecules with one (Figure 4A) or more (Figure 4B) 5' overhangs.

15

Figure 5 presents a process that allows ligation of blunt-ended molecules.

Figure 6 shows members of the tissue plasminogen activator gene family.

Figure 7 presents a list of certain polyketide compounds that are currently used as pharmaceutical drugs for the treatment of human and animal disorders.

Figure 8 depicts the different functional domains of bacterial polyketide synthase genes responsible for the production of erythromycin and rapamycin.

20

Figure 9 depicts the different functional domains of bacterial polyketide synthase genes responsible for the production of erythromycin and rapamycin.

Figure 10 depicts the protein functional domains of certain modular polyketide synthase genes.

25

Figure 11 presents a list of products generated by peptide synthetases that are currently used as pharmacologic agents.

Figure 12 depicts the protein functional domains of certain modular peptide synthetase genes.

Figure 13 depicts the structure of the *srfA* peptide synthetase operon.

30

Figure 14 depicts the synthesis of isoprenoids through the polymerization of isoprene building blocks.

Figure 15 depicts certain cyclization and intermolecular bond formation reactions catalyzed by isoprenoid, or terpene synthases.

Figure 16 presents a schematic illustration of the correspondence between natural exons and functional domains within isoprenoid synthases.

5 Figure 17 depicts one generic example of a directional ligation reaction.

Figure 18 presents a schematic representation of an inventive specific directional ligation reaction.

Figure 19A depicts the nucleotide sequence of the glutamate receptor exons known as Flip (GenBank accession number X64829).

10 Figure 19B depicts the nucleotide sequence of the glutamate receptor exons utilized are known as Flop (GenBank accession number X64830).

Figure 20 shows the amplified hybrid molecules produced in an inventive directional ligation reaction.

15 Figure 21 presents the nucleotide sequence of the ligation junction in the hybrid molecules of Figure 20.

Figure 22 presents the nucleotide sequence of the human β -globin gene.

Figure 23 shows an inventive identity exon shuffling reaction.

Figure 24 shows an inventive positional exon shuffling reaction.

20 Figure 25 shows the combinatorial potential of certain inventive directed ligation techniques.

Figure 26 presents one version of a combined primer-based/ribozyme-mediated nucleic acid manipulation scheme according to the present invention.

Figure 27 depicts a robotic system that could be utilized in the practice of certain inventive methods.

25 Figure 28 depicts a schematic representation of a directional ligation reaction employing inventive product molecules containing 3' overhangs.

Figure 29 presents a schematic of certain bioassay techniques that can be employed to determine the success of primer copying and/or ligation in inventive reactions.

30 Figure 30 shows a ribozyme mediated directional ligation reaction.

Figure 31 shows constructs employed in the reaction of Figure 30.

Figures 32 and 33 show products of the reaction of Figure 30.

Figures 34 shows a variety of chimeras generated using DNA-Overhang Cloning ("DOC"). The parental genes are shown in lines 1 and 2. The five chimeric genes are shown below the parental genes. Jagged edges indicate that only a portion of introns 13 and 15 were amplified. Lengths of chimeric genes (in basepairs) are indicated.

Definitions

"Cloning"-- The term "cloning", when used herein, means the production of a new nucleic acid molecule through the ligation of previously unlinked nucleic acid pieces to one another. A molecule produced by such ligation is considered a "clone" for the purposes of the present application, even before it has been replicated.

"Direct ligation"-- The term "direct ligation", as applied to product molecules herein, means that a product molecule may be ligated to one or more recipient molecules without first being cleaved with a restriction enzyme.

Preferably, no processing of the product molecule is required at all prior to ligation.

"Expression"-- "Expression" of nucleic acid sequences, as that term is used herein, means that one or more of (i) production of an RNA template from a DNA sequence; (ii) processing (e.g., splicing and/or 3' end formation) of a pre-mRNA to produce an mRNA; and (iii) translation of an mRNA has occurred.

"Gene"-- For the purposes of the present invention, the term "gene" has its art understood meaning. However, it will be appreciated by those of ordinary skill in the art that the term "gene" has a variety of meanings in the art, some of which include gene regulatory sequences (e.g., promoters, enhancers, etc.) and/or intron sequences, and others of which are limited to coding sequences. It will further be appreciated that art definitions of "gene" include references to nucleic acids that do not encode proteins but rather encode functional RNA molecules, such as tRNAs. For the purpose clarity, we note that, as used in the present application, the term "gene" generally refers to a portion of a nucleic acid that encodes a protein; the term may optionally encompass regulatory sequences. This definition is not intended to exclude application of the term "gene" to non-protein-coding expression units, but rather to clarify that, in most cases, the term as used in this document happens to be applied to a protein-coding nucleic acid.

“Gene fragment”-- A “gene fragment”, as that term is used herein, means a piece of a protein-coding DNA molecule that is shorter than the complete protein-coding molecule. Preferably, the fragment is at least about 12 bases long, more preferably at least about 15-20 bases long, and may be several hundred or thousands of base pairs long. It should be understood that the fragment need not include protein-coding sequence, but rather may represent a non-coding portion of the original gene.

“Hybrid nucleic acid”-- A “hybrid nucleic acid”, as that term is used herein, means a nucleic acid molecule comprising at least a first segment and a second segment, each of which occurs in nature but is not linked directly with the other in nature, the first and second segments being directly linked to one another in the hybrid nucleic acid.

“Overhang sequence”-- An “overhang sequence”, as that term is used herein, means a single stranded region of nucleic acid extending from a double stranded region.

“Primer”-- The term “primer”, as used herein, refers to a polynucleotide molecule that is characterized by an ability to be extended against a template nucleic acid molecule, so that a polynucleotide molecule whose sequence is complementary to that of at least a portion of the template molecule, is linked to the primer. Preferred primers are at least approximately 15 nt long. Particularly preferred primers have a length within the range of about 18-30, preferably longer than approximately 20 nucleotides

“Product molecule”-- A “product molecule”, as that term is used herein, is a nucleic acid molecule produced as described herein. Preferably, the product molecule is produced by extension of an oligonucleotide primer according to the present invention. A product molecule may be single stranded or double stranded. In certain preferred embodiments of the invention, a product molecule that includes a double-stranded portion also includes a single-stranded 3'- or 5'-overhang. In other preferred embodiments, the product molecule is blunt-ended. Where a product molecule is produced in an iterative DNA synthesis reaction (e.g., a PCR reaction), it is referred to as an “amplified product”.

“Recipient molecule”-- A “recipient molecule”, as that term is used herein, is a nucleic acid molecule to which a product molecule is to be ligated. The recipient molecule may be, but is not required to be, a vector. In general, the recipient molecule can be any molecule selected by the practitioner.

5 “Vector”-- A “vector”, as that term is used herein, is a nucleic acid molecule that includes sequences sufficient to direct *in vivo* or *in vitro* replication of the molecule. Where the vector includes *in vivo* replication sequences, these sequences may be self-replication sequences, or sequences sufficient to direct integration of the vector into another nucleic acid already present in the cell, so that
10 the vector sequences are replicated during replication of the already-present nucleic acid. Such already-present nucleic acid may be endogenous to the cell, or may have been introduced into the cell through experimental manipulation. Preferred vectors include a cloning site, at which foreign nucleic acid molecules, preferably inventive product molecules, may be introduced and ligated to the vectors.
15 Particularly preferred vectors further include control sequences selected for their ability to direct *in vivo* or *in vitro* expression of nucleic acid sequences introduced into the vector. Such control sequences may include, for example, transcriptional control sequences (e.g., one or more promoters, regulator binding sites, enhancers, terminators, etc.), splicing control sequences (e.g., one or more splice donor sites,
20 splice acceptor sites, splicing enhancers, etc.), and translational control sequences (e.g., a Shine Dalgarno sequence, a start codon, a termination codon, etc.). Vectors may also include some coding sequence, so that transcription and translation of sequences introduced into the vector results in production of a fusion protein.

25 Description of Certain Preferred Embodiments

Product molecules with 3' overhangs

In one aspect, the present invention provides reagents and methods for generating product molecules with 3' overhangs that can be directly ligated to recipient molecules. The length and sequence of the 3' overhang may be
30 determined by the user.

Figure 1 depicts one embodiment of this aspect of the invention. As shown in that Figure, first and second primers are provided that flank a target region of a

template nucleic acid molecule. At least one of the primers includes one or more ribonucleotides at its 5' end. Specifically, if primer 1 is x nucleotides long and primer 2 is y nucleotides long, then $n1 =$ a whole number (including 0) from 0 to x and $n2 =$ a whole number (including 0) from 0 to y except that (i) $n1$ and $n2$ cannot both be 0; and (ii) $n1$ can only be x (or $n2$ can only be y) if the DNA polymerase employed in the extension reaction is capable of extending an RNA primer. The characteristics (e.g., ability to extend an RNA primer, ability to copy RNA into DNA [whether the RNA is presented alone or as part of a hybrid RNA/DNA molecule]) of a wide variety of DNA polymerases are well known in the art (see, for example, manufacturer's catalogs, Myers et al., *Biochem.* 6:7661, 1991), and where such characteristics are not known for a particular DNA polymerase, routine assays are available for determining them (see, for example, Bebenek et al., *Met. Enzymol.* 262:217, 1995; see also Example 3).

In certain preferred embodiments of the invention, each of primers 1 and 2 includes at least one 5'-terminal ribonucleotide residue. In other preferred embodiments, at least one primer includes at least 2 ribonucleotide residues, one of which is the 5'-terminal residue. The primer may include at least 3, 4, 5, 6-10, or more ribonucleotide residues and even, as mentioned above, may be entirely RNA. Preferably, the ribonucleotide residues are contiguous with one another.

The nucleotide sequence of each of primer 1 and primer 2 is selected by the practitioner and need not be fully complementary with the sequence of the target nucleic acid. As is known in the art, perfect complementarity is not required for successful DNA synthesis, though it is generally desired that at least the 3'-terminal nucleotide of the primer be perfectly paired with the template. The 5' end of the primer, however, need not be paired at all, and it is common in the art to add additional sequences to a target sequence by including them in the primer. Of course, it is also acceptable for the primer to include a portion, 5' of the extendible 3' terminus, that does not hybridize with the template, and also to include a yet more 5' portion that does hybridize with the template. For the purposes of the present invention, any such variation on primer sequence, or any other available variation, is acceptable, so long as (i) at least one primer includes a ribonucleotide that either is present at the 5' end of the primer or will generate a new 5' end of

the primer upon being removed from the primer (e.g., by alkaline treatment, preferably followed by kinase treatment); and (ii) each primer hybridizes sufficiently well and sufficiently specifically under the conditions of the reaction that a product molecule is produced.

5 Other considerations of primer design are well known in the art (see, for example, Newton et al. (eds), *PCR: Essential Data Series*, John Wiley & Sons, New York, New York, 1995; Dieffenbach (ed), *PCR Primer: a Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1995; White et al. (eds), *PCR Protocols: Current Methods and Applications; Methods in Molecular*
10 *Biology*, The Humana Press, Totowa, NJ, 1993; Innis et al., *PCR Protocols: A Guide to Methods and Applications*, Academic Press, San Diego, CA, 1990; Griffin et al. (eds), *PCR Technology, Current Innovations*, CRC Press, Boca Raton, FL 1994, each of which is incorporated herein by reference). For instance, it is often desirable for approximately 50% of the hybridizing residues to be Gs or Cs; and
15 may be desirable, for optimal extension, for the 3'-terminal residue to also be a G or a C.

Primers such as those depicted in Figure 1, that contain at least one ribonucleotide residue as their 5' terminal residue (or as a residue whose removal will create a new 5'-terminal primer residue), may be prepared by any technique
20 available in the art. For example, such primers may be chemically synthesized. Companies (e.g., Oligos, Etc., Inc., Bethel, ME) that supply oligonucleotide reagents will typically prepare hybrid RNA/DNA oligonucleotides, or RNA only nucleotides, as preferred by the practitioner. Alternatively, RNA sequences may be ligated to DNA sequences using standard techniques (see, for example, Moore et
25 al., *Science* 256:992, 1992; Smith (ed), *RNA: Protein Interactions, A Practical Approach*, Oxford University Press, 1998, which particularly discusses construction of RNA molecules containing site-specific modifications by RNA ligation; each of these references is incorporated herein by reference).

As shown in Figure 1, an extension reaction is performed so that DNA
30 synthesis is primed from each of the first and second primers, and a double stranded DNA/RNA hybrid molecule is created with at least one ribonucleotide residue at the 5' end of at least one strand. Preferably, but not essentially, the

DNA polymerase utilized in the extension reaction is one that does not add extraneous 3' nucleotides. Also, as mentioned above, if one or both of the primers has a ribonucleotide as its 3' residue, the DNA polymerase utilized in the extension step must be one that is capable of extending from a ribonucleotide primer.

5 Figure 1 shows that the hybrid molecule is then exposed to a treatment that removes the ribonucleotide residues. As depicted in Figure 1, that treatment is exposure to elevated pH (e.g., treatment with a base such as sodium hydroxide [NaOH]). Any other treatment that removes RNA residues without disturbing DNA residues (e.g., exposure to RNase, etc.) could alternatively be employed at
10 this step.

 When the ribonucleotide residues are removed from the hybrid molecule, the resultant molecule is left with a double stranded portion and a single stranded 3' overhang on at least one of its ends. Figure 1 depicts a product molecule with single-stranded 3' overhangs at both ends. The sequence and length of the
15 overhang was determined by the sequence and length of RNA present at the 5' end of the primers. Clearly, any sequence and length of overhang can be selected. In certain preferred embodiments of the invention, the sequence and length of the overhang corresponds with that produced by cleavage of double-stranded DNA by a commercially available restriction enzyme, so that the product molecule can be
20 ligated to recipient molecules that have been cut with that enzyme. A variety of enzymes that leave 3' overhangs are known in the art, including but not limited to *AatII*, *AlwI*, *NsiI*, *SphI*, etc.

 In other preferred embodiments, the 3' overhang sequence and length is selected to base pair with a 3' overhang generated in another inventive product
25 molecule, so that the two molecules may readily be ligated together (see, for example, Example 1).

 Furthermore, it will be appreciated that the 3' overhangs at the two ends of the product molecule need not have the same sequence or length (see, for example, Example 1). It is often desirable to generate a nucleic acid molecule that can be
30 ligated to a recipient molecule in only one orientation, or that can be ligated to two different recipient nucleic acid molecules (e.g., a three-way ligation) in a particular

arrangement. Accordingly, it is quite valuable to be able to engineer the sequence and length of the 3' overhangs of the inventive product molecule.

As can be seen with reference to Figure 1, the nature of the ends left by the ribonucleotide-removal treatment can affect the behavior of the product molecule in subsequent ligation reactions. In particular, alkaline hydrolysis of ribonucleotides leaves 5' -OH groups rather than 5'-phosphate groups. As is known in the art, at least one terminal phosphate group is typically required for successful ligation of nucleic acid molecules. Thus, if the product molecule depicted in Figure 1 is to be ligated to a recipient molecule that lacks the appropriate terminal phosphate groups (e.g., because of exposure to treatment with a phosphatase), it will be desirable to add 5' phosphate groups to the recipient molecule prior to ligation. Any available technique may be utilized to achieve such phosphate group addition; most commonly, the phosphate groups will be added by treatment with polynucleotide kinase.

The product molecules depicted in Figure 1 may be ligated to any desired recipient molecule. Preferably, the recipient molecule has at least one 3' overhang that is complementary to at least a portion of the at least one 3' overhang on the product molecule. It will be appreciated that, if the recipient molecule has a 3' overhang whose 3' terminal portion is complementary to the 3' terminal portion of the product molecule 3' overhang, but is not otherwise complementary to the product molecule 3' overhang, then one or more gaps will be present after hybridization, which gaps can be filled in with DNA polymerase prior to ligation. Since the sequence and length of the product molecule 3' overhang is selected by the practitioner, this approach may be employed to add sequence to the recombinant molecule that would not be present if complete 3' overhang hybridization had occurred. For the purposes of the present invention, the complementary 3'-terminal portions of the product and recipient molecules should be at least one nucleotide long, and can be 2, 3, 4, 5, 6-10 nucleotides long, or longer. In certain preferred embodiments, the complementary 3'-terminal portions are less than about 6 nucleotides long, so that efficiency of ligation (usually performed at 4 °C or 14 °C) is preserved and complications associated with annealing longer sequences are avoided.

Preferred recipient molecules include, but are not limited to, linearized vectors. Such vectors may be linearized, for example by digestion with a restriction enzyme that produces a 3' overhang complementary to that present on the product molecule. Alternatively, such linearized vectors may be prepared as product molecules as described herein, containing one or more 3' overhangs selected by the practitioner to be compatible with the 3' overhangs present on other product molecules.

Those of ordinary skill in the art will appreciate that product molecules can readily be generated according to the present invention so that each end of a given product molecule has a different 3' overhang. Such molecules can be used in directional cloning experiments, where they can be ligated to one or more other molecules in only a single orientation. Such directional ligation strategies are particularly useful where three or more molecules are desired to be linked to one another. In such multi-component ligation reactions, it is often useful to minimize the possibility of self-ligation by individual molecules, and also to reduce the chance that the molecules will link together with one or more molecule being in an improper orientation.

Product molecules with 5' overhangs

Figures 2 - 4 depict inventive strategies for producing product molecules with 5' overhangs. For example, as shown in Figure 2, a template molecule may be hybridized with one or more primers including at least one ribonucleotide. For this embodiment of the present invention, it is not required that the ribonucleotide be located at the 5' end of the oligonucleotide, though such is acceptable. The primer may contain 2, 3, 4, 5, 6-10, or more ribonucleotides, and may be wholly ribonucleotides if the DNA polymerase utilized in the extension reaction will extend a ribonucleotide primer. That is, in Figure 2, at least one of n_1 and n_2 is a whole number greater than or equal to 1, and n_3 and n_4 are each a whole number greater than or equal to zero. The particular inventive embodiment depicted in Figure 3 utilizes two primers. Those of ordinary skill in the art will appreciate that each primer includes a portion, terminating with the 3'-terminal residue of the primer, that hybridizes sufficiently well with the template molecule to allow

extension. The sequence of the remainder of the primer, however, need not be complementary to that of the template molecule. Furthermore, those of ordinary skill in the art will also appreciate that if the DNA polymerase being employed includes a 3' → 5' exonuclease activity, it is not even essential that the 3'-most residue in the primer hybridize with the template, so long as the exonuclease activity is allowed to chew back to a point in the primer from which extension can occur.

After hybridization with the primer(s), an extension reaction is performed with a DNA polymerase that does not copy ribonucleotides. For example, we have found that Vent_R[®] and Vent_R[®] (exo⁻) do not use ribonucleotide bases as a template (see Example 1); *Tth* and *Taq* polymerases, by contrast, are reported to be able to replicate ribonucleotides (Myers et al., *Biochem.* 6:7661, 1991), as, of course, are reverse transcriptases. Other DNA polymerases may be tested for their ability to copy ribonucleotides according to standard techniques or, for example, as described in Example 3.

The extension reaction shown in Figure 2 may be iterated as an amplification reaction, if desired. The embodiment depicted in Figure 2 illustrates such an amplification, from which the product is a double stranded molecule with two 5' overhangs, each of which includes at least one ribonucleotide residue. Those of ordinary skill in the art will appreciate that the sequence and length of each 5' overhang (as well as is ribonucleotide composition) is selected by the practitioner, and that the two product molecule overhangs depicted may be the same or different.

This product molecule may then be hybridized with one or more recipient molecules containing a 5' overhang that is complementary to at least the 5'-terminal residue of the product molecule. If gaps remain after hybridization, they may be filled in with DNA polymerase according to known techniques. If the gaps encompass a ribonucleotide residue, it may be preferable to employ a DNA polymerase that will copy RNA in order to ensure that the gap is filled. As mentioned above, such DNA polymerases include, for example, *Tth*, *Taq*, and reverse transcriptase. Other DNA polymerases may be tested for their ability to

copy RNA according to known techniques or, for example, as described in Example 3.

Once any gaps are filled, the product and recipient molecules may be ligated together. DNA ligase is known to close nicks (i) between adjacent
5 deoxyribonucleotides; (ii) between a deoxyribonucleotide and a ribonucleotide; or (iii) between adjacent ribonucleotides. Thus, a hybrid molecule can be produced containing both DNA and RNA residues. This molecule can be copied into DNA, either *in vitro* according to standard techniques, or *in vivo* after introduction in to a host cell capable of copying such a molecule (*Escherichia coli*, for example, have
10 been reported to be able to remove and replace ribonucleotides that are base-paired with deoxyribonucleotides-- see Sancar, *Science* 266:1954, 1994). Alternatively, it may be desirable to replicate the hybridized compound into DNA rather than performing a ligation (e.g., by PCR with DNA primers or with a DNA polymerase that can copy ribonucleotides). Also, it should be mentioned that, in some cases,
15 ligation may be accomplished *in vivo* rather than *in vitro*, as is known in the art for example for co-transformation of yeast cells.

As depicted in Figure 2, the product molecule is ligated with only a single recipient molecule and at only one end. Those of ordinary skill in the art will appreciate that a product molecule may alternatively be ligated at both of its ends,
20 either to a single recipient molecule or to two different recipient molecules.

Figure 3 presents an alternative approach to generating product molecules with one or more 5' overhangs. In this embodiment, instead of employing ribonucleotide primer residues and a DNA polymerase that cannot copy RNA, we utilize a modified base in the primer, which modified base is not copied by the
25 DNA polymerase. A wide variety of modified nucleotides are known in the art (see, for example, U.S. Patent Number 5,660,985; see also various catalogs such as that provided by Oligos, Etc. [Bethel, ME]); those that are not copied by particular DNA polymerases may be identified, for example, by reference to the manufacturer's catalog, by routine screening according to known techniques, or as
30 described, for example, in Example 3.

Modified bases may be removed from the product molecule, before or after its ligation to a recipient molecule, either by DNA replication *in vitro* or *in vivo*

with a DNA polymerase that will copy the modified base or by removal of the base followed by gap repair, according to standard techniques (see, for example, Sancar, *Science* 266:1954, 1994).

Figure 4 presents an inventive embodiment for generating a product molecule with at least one 5' overhang. In the particular embodiment depicted in Figure 4, the inventive strategy is applied to a starting molecule containing one (Figure 4A) or two (Figure 4B) 3' overhangs, so that the starting molecule is converted from a 3'-overhang-containing compound to a 5'-overhang-containing molecule. However, those of ordinary skill in the art will appreciate that the same approach could equally well be applied to add one or two 5' overhangs to a starting molecule that is either blunt ended, or contains one or two 3' or 5' overhangs.

The starting molecule depicted in Figure 4 may be obtained by any available means. The molecule may have one or two 3' overhangs (meaning that at least one of R1 and R2 is at least one nucleotide long) and may be produced, for example, by restriction endonuclease cleavage of a precursor fragment, by polymerase chain amplification, or by any other means. In certain preferred embodiments of the invention, the starting molecule is produced by PCR and contains a single 3' dATP at each end, as described above. Figure 4A depicts the application of the inventive method to a starting molecule having only one 3' overhang; Figure 4B depicts the application of the inventive method to a starting molecule having two 3' overhangs.

With reference to Figure 4A, the starting molecule is hybridized with at least one primer containing a first portion that hybridizes with a first sequence in the starting molecule that is substantially adjacent to the starting molecule 3' overhang residue, a second portion that aligns with and fails to hybridize to at least one residue of the starting molecule 3' overhang, and a third portion that does not align with the starting molecule but rather extends past (5' in the primer) the last residue of the starting molecule 3' overhang.

The length and sequence of the first portion of the primer is determined by the sequence of the starting molecule adjacent the starting molecule 3' overhang. Hybridization by the first portion of the primer may extend into the 3' overhang, so long as at least one residue of the 3' overhang is aligned with and fails to hybridize

with the second portion of the primer. The length and sequence of the second portion of the primer is determined to some degree by the length and sequence of the starting molecule 3' overhang in that the second portion must fail to hybridize with at least one residue of the 3' overhang, preferably but not essentially at least the 3'-terminal residue. So long as such hybridization is avoided, the precise sequence of this second portion of the primer may be selected by the practitioner. The length (i.e., the value of n in Figure 4A, which must be a whole number greater than or equal to 1) and sequence (i.e., the identities of N in Figure 4A) of the third portion of the primer is also determined by the practitioner. This third portion will become a 5' overhang in the product molecule.

As depicted in Figure 4A, single or multiple rounds of extension from the inventive primer is performed. It will be appreciated by those of ordinary skill in the art that, due to the absence of a second primer (and the mismatch between the primer and the starting molecule 3' overhang, which prevents extension of the 3' end of that strand of the starting molecule) only linear, and not exponential, extension is accomplished. Of course, if the DNA polymerase employed in the extension reaction is one that adds one or more terminal 3' residues, the product molecule may have a 3' overhang as well as a 5' overhang.

Once the product molecule with a 5' overhang is produced, it may be hybridized with any recipient molecule that also contains a 5' overhang, at least part of which is complementary to part of the product molecule 5' overhang. The hybridized compound contains a nick on each strand (or may even contain a gap if the 5' overhangs of the product and recipient molecules are imperfectly matched in length) and at least one mismatch immediately prior to the product molecule 5' overhang. This hybridized compound is then exposed to a 3'→5' exonuclease activity to remove the mismatched base(s) (that correspond to the portion of the starting molecule 3' overhang that did not hybridize with the second portion of the primer). The digested compound is then exposed to a DNA polymerase to fill in the gap created by exonuclease digestion, and subsequently to ligase to heal any remaining nicks. Enzymes having 3' → 5' exonuclease activity are well known in the art (including, for example, *E. coli* DNA polymerase I, *Pfu*, Vent_R[®], Deep

Vent_R[®], etc.); other enzymes may be tested for this ability according to standard techniques.

Those of ordinary skill in the art will appreciate that the method depicted in Figure 4A may be applied to either strand of a starting molecule, depending on where the 3' overhang is located. As depicted in Figure 4B, the method may even be applied to both strands simultaneously, although it is important for such an embodiment to perform only a single round extension reaction or to perform independent extension reactions for each strand. Amplification (i.e., multiple rounds of denaturation and extension) is not performed because such amplification would result in the production of a blunt-ended molecule (or one with 3' overhangs if a DNA polymerase that adds 3' nucleotides were employed), having the sequence dictated by the primers, rather than a molecule with a 5' overhang and a mismatch immediately 3' of the 5' overhang.

As shown in Figure 4B, a starting molecule containing two 3' overhangs is converted to a product molecule containing two 5' overhangs by application of the inventive method. The starting molecule is hybridized with two inventive primers containing first, second, and third portions as described above in the discussion of Figure 4A. Each primer is then extended in single-round (or independent) extension reactions. It will be understood by those of ordinary skill in the art that both extension reactions need not be performed simultaneously, or on the same exact starting molecule. Extensions of each primer can even be performed in different reaction vessels.

Each of the double-stranded molecules produced in the extension reaction has a single 5' overhang, whose sequence and length corresponds to that of the third primer portion. The strands of these double stranded molecules are then separated from one another. Individual strands may be separately purified if desired, but such is not required. Strands are then mixed together (if they are not already together) and annealed, so that the two new strands synthesized by extension of the primers have the opportunity to anneal to one another. The product of this annealing reaction is an inventive product molecule with two 5' overhangs. As will be appreciated, these overhangs may be the same or different in length and/or sequence.

This product molecule may be hybridized with one or more recipient molecules, each of which has a 5' overhang whose 5'-terminal portion (at least one nucleotide in length) is complementary with a 5'-terminal portion (of the same length) of the product molecule 5' overhang. Any gaps remaining after
5 hybridization may be filled in with a DNA polymerase; the product and recipient molecules may then be ligated together.

Blunt-ended product molecules

Figure 5 presents an inventive embodiment that allows ligation of blunt-
10 ended molecules. As shown, blunt ended starting molecules are provided that are to be linked together. Such molecules may be prepared by any available technique including, for example, digestion of a precursor with one or more restriction enzymes (optionally followed by a fill-in or chew-back of any overhanging ends), PCR (e.g., with a DNA polymerase that does not add extraneous 3' nucleotides--
15 reference can be made to manufacturer's catalogs to determine the characteristics of a particular DNA polymerase. For example, Vent_R[®] is reported to generate > 95% blunt ends; Vent_R[®] (exo⁻) is reported to generate about 70% blunt ends and 30% single nucleotide 3' overhangs, of any nucleotide; Pfu is reported to produce only blunt-ended molecules), chemical synthesis, etc. The starting molecules may be
20 double stranded or single stranded. As depicted in Figure 5, the starting molecules are double stranded.

The starting molecules are hybridized to bridging molecules, each of which hybridizes to at least one terminal residue of two different starting molecules that are to be linked together. Clearly, if the starting molecules are double stranded,
25 they should be denatured prior to exposure to the bridging molecules, so that successful hybridization with the bridging molecules may occur. The bridging molecules may hybridize to more than one residue of each starting molecule, and/or may contain non-hybridizing portions between the portions that hybridize to the two starting molecules. Also, the bridging molecules may have sufficient length
30 that they abut one another after hybridization, or may be short enough that gaps are present in the hybridized compound between the individual bridging molecules. Preferably, at least one primer hybridizes to the 3'-terminus of the 3'-most starting

molecule in the hybridized compound. This primer may extend past the terminus if desired, so that a 5' overhang is created. No such overhang is depicted in Figure 5.

5 The hybridized compound is then converted into a double-stranded DNA molecule by any collection of available techniques. For example, gaps may be filled with DNA polymerase and any remaining nicks sealed with DNA ligase. Or, if no gaps are present in one strand, that strand may first be ligated and DNA polymerase subsequently applied, *in vitro* or *in vivo* to seal gaps in the other strand or to synthesize a replacement strand (e.g., primed from the bridging molecule
10 hybridized at the most 3' location with respect to the starting molecules). In one preferred embodiment of the invention, gaps are filled and nicks sealed and the entire recombinant molecule is then replicated by PCR amplification. If desired, a DNA polymerase that adds one or more 3'-terminal residues may be employed, so that the resultant amplified product is likely to have one or more 3' overhangs. As
15 described above, such a product may readily be ligated to another molecule with complementary 3' overhangs, such as occurs in the use of the Invitrogen TA Cloning Kit® system.

Applications

20 The product molecules and ligation strategies provided above are useful in any of a variety of contexts. For the purposes of clarification only, and not for limitation, we discuss certain of these contexts in more detail here.

As described above, the present invention produced techniques and reagents for providing nucleic acid molecules that can be directly ligated (i.e., without first
25 being digested with a restriction enzyme) to other molecules. The invention also provides techniques for accomplishing such ligation. The present invention may be used to link nucleic acid molecules of any sequence to one another and therefore has the broadest possible application in the field of genetic cloning.

Those of ordinary skill in the art will appreciate that the inventive
30 techniques and reagents may be employed to link any DNA molecule to any other DNA molecule, regardless of the particular sequences of the DNA molecules, their protein-coding capacities, or any other characteristics. This feature distinguishes

the present system from traditional, restriction-endonuclease-reliant cloning systems, for which the precise sequences of the molecules being linked can often affect the design of the cloning strategy, as it may be desirable, for example, to avoid cleaving one fragment with a particular enzyme that produces an undesired
5 cleavage in another fragment, or to make other adjustments to accommodate the behavior of the protein enzymes being employed.

Production of protein-coding genes

In certain preferred embodiments of the present invention, one or more of
10 the DNA molecules included in an inventive ligation reaction includes open reading frame, i.e., a protein-coding sequence. In particularly preferred embodiments, at least two DNA molecules to be ligated together include open reading frame sequences, and their ligation produces a hybrid DNA containing both open reading frames linked together so that a single polypeptide is encoded. Where ligation of
15 two or more DNA molecules, according to the present invention, generates at least one open reading frame that spans at least one ligation junction, the ligation is considered to have generated a new, hybrid protein-coding gene.

In but one embodiment of the inventive system used to produce protein-coding genes, the DNA molecules to be ligated to one another are selected to
20 encode one or more discrete functional domains of known biological activity, so that the ligation of two or more such DNA molecules produces a hybrid gene encoding a bi- or multi-functional polypeptide. It is well known in the art that many proteins have discrete functional domains (see, for example, Traut, *Mol. Cell. Biochem.* 70:3, 1986; Go et al., *Adv. Biophys.* 19:91, 1985). It is also well known
25 that such domains may often be separated from one another and ligated with other discrete functional domains in a manner that preserves the activity of each individual functional domain.

Those of ordinary skill in the art will appreciate that some flexibility is allowed in the selection of precise DNA sequences encoding functional protein
30 domains. For example, it is often not desirable to limit the DNA sequences to only those that encode for exactly the amino acid residues contained in a functional domain of a naturally-occurring protein. Additional DNA sequences may be

included, for example, encoding linker sequences that can provide flexibility between the particular selected functional domain and any other functional domain to which it is to be linked.

Alternatively or additionally, in some contexts researchers have found that it is useful to select DNA sequences encoding less than all of the amino acids comprising a particular functional domain (see, for example, WO 98/01546); in such cases, the other amino acids can be added back as a result of the subsequent ligation (i.e., can be encoded by an adjacently-ligated DNA molecule), or can be left out completely. Those of ordinary skill in the art will readily be able to familiarize themselves with the application of these basic principles to their particular experimental question after appropriate consultation with the literature describing the protein domains in which they are interested.

To give but a few examples of the types of functional protein domains that could be encoded by individual DNA molecules, or combinations of DNA molecules, to be ligated according to the present invention, well known modular domains include, for example DNA binding domains (such as zinc fingers, homeodomains, helix-turn-helix motifs, etc.), ATP or GTP binding domains, transmembrane spanning domains, protein-protein interaction domains (such as leucine sippers, TPR repeats, WD repeats, STYX domains [see, for example, Wishart et al., *Trends Biochem. Sci.* 23:301, 1998], etc.), G-protein domains, tyrosine kinase domains (see, for example, Shokat, *Chem. Biol.* 2:509, 1995), SRC homology domains (see, for example, Sudol, *Oncogene* 17:1469, 1998), SH2 domains (see, for example, Schaffhausen, *Biochim. Biophys. Acta* 28:61, 1995), PTB domains (see, for example, van der Greer et al., *Trends Biochem Sci* 20:277, 1995), the PH domain (see, for example, Musacchio et al., *Trends Biochem Sci* 18:343, 1993), certain catalytic domains, cell surface receptor domains (see, for example, Campbell et al., *Immunol. Rev.* 163:11, 1998), carbohydrate recognition domains (see, for example, Kishore et al., *Matrix Biol.* 15:583, 1997), immunoglobulin domains (see, for example, Rapley, *Mol. Biotechnol.* 3:139, 1995), etc. (see also, Hegyi et al., *J. Protein. Chem.* 16:545, 1997; Baron et al., *Trends Biochem. Sci.* 16:13, 1997).

Typically, such domains are identified by homology comparisons that identify regions of sequence similarity within proteins of known biological activity (at least as relates to the portion of the protein showing the homology). The spatial coherence of any particular functional domain is often confirmed by structural studies such as X-ray crystallography, NMR, etc.

According to the present invention, a useful "functional domain" of a protein is any portion of that protein that has a known biological activity that is preserved with the portion is separated from the rest of the protein, even if the portion must continue to be embedded within a larger polypeptide molecule in order to maintain its activity. The relevant biological activity need not, and typically will not, constitute the complete biological activity of a particular protein in which the domain is naturally found, but rather will usually represent only a portion of that activity (e.g., will represent an ability to bind to a particular other molecule but will not include a further activity to cleave or modify the bound molecule). As noted, many such domains have already been described in the literature; others can be identified by homology search, preferably in combination with mutational studies as is known in the art to define sequences that participate in biological activity.

The present invention encompasses the recognition, now virtually universally accepted, that the production of new genes during evolution has often involved the novel combination of DNA sequences encoding two or more already-existing functional protein domains (see, for example, Gilbert et al., *Proc Natl Acad Sci USA*, 94:7698, 1997; Strelets, et al., *Biosystems*, 36:37, 1995). In fact, protein "families" are often defined by their common employment of particular functional domains, even though the overall biological roles played by different family members may be quite unrelated (see further discussion of such families below, in section discussing exon shuffling). The present invention therefore provides techniques and reagents that can be used to mimic an evolutionary process in the laboratory. The universality and experimental simplicity of the system provide researchers, who may select particular DNA modules to link to one another in desired orders, with significant advantages over Mother Nature, who must wait for stochastic processes to produce interesting new results.

Accordingly, preferred protein functional domains to be employed in accordance with the present invention include those that have been re-used through evolution to generate gene families (i.e., collections of genes that encode different members of protein families). Exemplary gene families created by re-use of particular protein domains include, for example, the tissue plasminogen activator gene family (see, for example Figure 6); the family of voltage-gated sodium channels (see, for example, Marban et al., *J. Physiol.* 508:647, 1998); certain families of adhesion molecules (see, for example, Taylor et al., *Curr. Top. Microbiol. Immunol.* 228:135, 1998); various extracellular domain protein families (see, for example, Engel, *Matrix Biol.* 15:295, 1996; Bork, *FEBS Lett.* 307:49, 1992; Engel, *Curr. Opin. Cell. Biol.* 3:779, 1991); the protein kinase C family (see, for example, Dekker et al., *Curr. Op. Struct. Biol.* 5:396, 1995); the tumor necrosis factor receptor superfamily (see, for example, Naismith et al., *J. Inflamm* 47:1, 1995); the lysin family (see, for example, Lopez et al., *Microb. Drug Resist.* 3:199, 1997); the nuclear hormone receptor gene superfamily (see, for example, Ribeiro et al., *Annu. Rev. Med.* 46:443, 1995; Carson-Jurica et al., *Endocr. Rev.* 11:201, 1990); the neurexin family (see, for example, Missler et al., *J. Neurochem.* 71:1339, 1998); the thioredoxin gene family (see, for example, Sahrawy et al., *J. Mol. Evol.*, 42:422, 1996); the phosphoryl transfer protein family (see, for example, Reizer et al., *Curr. Op. Struct. Biol.* 7:407, 1997); the cell wall hydrolase family (see, for example, Hazlewood et al., *Prog. Nuc. Acid Res. Mol. Biol.* 61:211, 1998); as well as certain families of synthetic proteins (e.g., fatty acid synthases, polyketide synthases [see, for example, WO 98/01546; U.S. Patent Number 5,252,474; U.S. Patent Number 5,098,837; EP Patent Application Number 791,655; EP Patent Application Number 791,656], peptide synthetases [see, for example, Mootz et al., *Curr. Op. Chem. Biol.* 1:543, 1997; Stachelhaus et al., *FEMS Microbiol. Lett* 125:3, 1995], and terpene synthases).

The present invention allows DNA molecules encoding different functional domains present in these families to be linked to one another to generate in-frame fusions, so that hybrid genes are produced that encode polypeptides containing different arrangements of the selected functional domains. It will be appreciated that experiments can be performed in which (i) only the domains utilized in a

particular gene family in nature are linked to one another (in new arrangements), or in which (ii) domains naturally utilized in different gene families are linked to one another.

In one particularly preferred embodiment of the present invention, the DNA
5 modules selected to be ligated together comprise modules encoding at least one functional domain, or portion of a functional domain, of a member of a synthetic enzyme family. As mentioned above, a variety of enzyme families are known whose members are responsible for the synthesis of related biologically active compounds. Families of particular interest include the fatty acid synthase family,
10 the polyketide synthase family, the peptide synthetase family, and the terpene synthase family (sometimes called the terpenoid synthase family, or the isoprenoid synthase family). The individual members of these enzyme families are multi-domain proteins that catalyze the synthesis of particular biologically active chemical compounds. For any particular family member, different protein domains
15 catalyze different steps in the overall synthesis reaction. Each family member catalyzes the synthesis of a different chemical compound because each contains a different collection or arrangement of protein functional domains. As will be understood in the context of the present application, the instant invention provides a system by which the various protein domains utilized in these gene families may be
20 linked to one another in new ways, to generate novel synthase enzymes that will catalyze the production of new chemical entities expected to have biological activities related to those produced by naturally-occurring members of the gene family from which the functional domains were selected.

In order to more clearly exemplify this aspect of the present invention, we
25 discuss below certain characteristics and attributes of each of the above-mentioned particularly preferred synthetic enzyme protein families:

ANIMAL FATTY ACID SYNTHASE FAMILY

The animal fatty acid synthase comprises two multifunctional polypeptide
30 chains, each of which contains seven discrete functional domains. Fatty acid molecules are synthesized at the interface between the two polypeptide chains, in a reaction that involves the iterative condensation of an acetyl moiety with successive

malonyl moieties (see, for example, Smith, *FASEB J.* 8:1248, 1994; Wakil, *Biochemistry* 28:4523, 1989, each of which is incorporated herein by reference).

Most commonly, the β -keto intermediate produced in this condensation reaction is completely reduced to produce palmitic acid; in certain instances, however,
5 alternative substrates or alternative chain-terminating mechanisms are employed so that a range of products, including branched-chain, odd carbon-numbered, and shorter-chain-length fatty acid molecules. These molecules have a range of roles in biological systems, including (i) acting as precursors in the production of a variety of signalling molecules, such as steroids, as well as (ii) participating in the
10 regulation of cholesterol metabolism.

Those of ordinary skill in the art, considering the present disclosure, will readily recognize that the techniques and reagents described herein can desirably be applied to DNA molecules encoding one or more of the functional domains of a fatty acid synthase molecule, so that the molecules may be linked to other DNA
15 molecules to create interesting new hybrid DNAs, preferably encoding hybrid animal fatty acid synthase genes that may have novel synthetic capabilities.

POLYKETIDE SYNTHASE FAMILY

Polyketides represent a large and structurally diverse class of natural
20 products that includes many important antibiotic, antifungal, anticancer, antihelminthic, and immunosuppressant compounds such as erythromycins, tetracyclines, amphotericins, daunorubicins, avermectins, and rapamycins. For example, Figure 7 presents a list of certain polyketide compounds that are currently used as pharmaceutical drugs in the treatment of human and animal disorders.

Polyketides are synthesized by protein enzymes, aptly named polyketide
25 synthases, that catalyze the repeated stepwise condensation of acylthioesters in a manner somewhat analogous to that employed by the fatty acid synthases. Structural diversity among polyketides is generated both through the selection of particular "starter" or "extender" units (usually acetate or propionate units)
30 employed in the condensation reactions, and through differing degrees of processing of the β -keto groups observed after condensation. For example, some β -keto groups are reduced to β -hydroxyacyl- groups; others are both reduced to this

point, and are subsequently dehydrated to 2-enoyl groups; still others are reduced all the way to the saturated acylthioester.

Polyketide synthases (PKSs) are modular proteins in which different functional domains catalyze different steps of the synthesis reactions (see, for example, Cortes et al., *Nature* 348:176, 1990; MacNeil et al., *Gene* 115:119, 1992; Schwecke et al., *Proc. Natl. Acad. Sci. USA* 92:7839, 1995). For example, Figures 8 and 9 (from WO 98/01546) depict the different functional domains of bacterial polyketide synthase genes responsible for the production of erythromycin and rapamycin, respectively (see also Figure 10). Each of these genes is an example of a so-called "class I" bacterial PKS gene. As shown, each cycle of polyketide chain extension is accomplished by a catalytic unit comprising a collection of functional domains including a β -ketoacyl ACP synthase domain (KS) at one end and an acyl carrier protein (ACP) domain at the other end, with one or more other functional domains (selected from the group consisting of an acyl transferase [AT] domain, a β -ketoacyl reductase [KR] domain, an enoyl reductase [ER] domain, a dehydratase [DH] domain, and a thioesterase [TE] domain).

Class II bacterial PKS genes are also modular, but encode only a single set of functional domains responsible for catalyzing chain extension to produce aromatic polyketides; these domains are re-used as appropriate in successive extension cycles (see, for example, Bibb et al., *EMBO J.* 8:2727, 1989; Sherman et al., *EMBO J.* 8:2717, 1989; Fernandez-Moreno et al., *J. Biol. Chem.* 267:19278, 1992; Hutchinson et al., *Annu. Rev. Microbiol.* 49:201, 1995). Diversity is generated primarily by the selection of particular extension units (usually acetate units) and the presence of specific cyclases (encoded by different genes) that catalyze the cyclization of the completed chain into an aromatic product.

It is known that various alterations in and substitutions of class I PKS functional domains can alter the chemical composition of the polyketide product produced by the synthetic enzyme (see, for example, Cortes et al., *Science* 268:1487, 1995; Kao et al., *J. Am. Chem. Soc.* 117:9105, 1995; Donadio et al., *Science* 252:675, 1991; WO 93/1363). For class II PKSs, it is known that introduction of a PKS gene from one microbial strain into a different microbial strain, in the context of a different class II PKS gene cluster (e.g., different

cyclases) can result in the production of novel polyketide compounds (see, for example, Bartel et al., *J. Bacteriol.* 172:4816, 1990; WO 95/08548).

The present invention provides a new system for generating altered PKS genes in which the arrangement and/or number of functional domains encoded by the altered gene differs from that found in any naturally-occurring PKS gene. Any

PKS gene fragment can be used in accordance with the present invention. Preferably, the fragment encodes a PKS functional domain that can be linked to at least one other PKS functional domain to generate a novel PKS enzyme. A variety of different polyketide synthase genes have been cloned (see, for example, Schwecke et al., *Proc. Natl. Acad. Sci. USA* 92:7839, 1995; U.S. Patent Number 5,252,474; U.S. Patent Number 5,098,837; EP Patent Application Number 791,655; EP Patent Application Number 791,656, each of which is incorporated herein by reference; see also WO 98/51695, WO 98/49315, and references cited therein, also incorporated by reference.), primarily from bacterial or fungal organisms that are prodigious producers of polyketides. Fragments of any such genes may be utilized in the practice of the present invention.

PEPTIDE SYNTHETASE FAMILY

Peptide synthetases are complexes of polypeptide enzymes that catalyze the non-ribosomal production of a variety of peptides (see, for example, Kleinkauf et al., *Annu. Rev. Microbiol.* 41:259, 1987; see also U.S. Patent Number 5,652,116; U.S. Patent Number 5,795,738). These complexes include one or more activation domains (DDA) that recognize specific amino acids and are responsible for catalyzing addition of the amino acid to the polypeptide chain. DDA that catalyze the addition of D-amino acids also have the ability to catalyze the racemization of L-amino acids to D-amino acids. The complexes also include a conserved thioesterase domain that terminates the growing amino acid chain and releases the product. Figure 11 presents an exemplary list of products generated by peptide synthetases that are currently being used as pharmacologic agents.

The genes that encode peptide synthetases have a modular structure that parallels the functional domain structure of the enzymes (see, for example, Cosmina et al., *Mol. Microbiol.* 8:821, 1993; Kratzschmar et al., *J. Bacteriol.* 171:5422,

1989; Weckermann et al., *Nuc. Acids res.* 16:11841, 1988; Smith et al., *EMBO J.* 9:741, 1990; Smith et al., *EMBO J.* 9:2743, 1990; MacCabe et al., *J. Biol. Chem.* 266:12646, 1991; Coque et al., *Mol. Microbiol.* 5:1125, 1991; Diez et al., *J. Biol. Chem.* 265:16358, 1990; see also Figure 12). For example, Figure 13 (from U.S. Patent Number 5,652,116) presents the structure of one exemplary peptide synthetase gene operon, the *srfA* operon.

The sequence of the peptide produced by a particular peptide synthetase is determined by the collection of functional domains present in the synthetase. The present invention, by providing a system that allows ready linkage of particular peptide synthetase functional domains to one another, therefore provides a mechanism by which new peptide synthase genes can be produced, in which the arrangement and/or number of functional domains is varied as compared with naturally-occurring peptide synthase genes. The peptide synthase enzymes encoded by such new genes are expected to produce new peptide products. The present invention therefore provides a system for the production of novel peptides, through the action of hybrid peptide synthase genes.

TERPENE SYNTHASE FAMILY

Isoprenoids are chemical compounds whose structure represents a modification of an isoprene building block. The isoprenoid family includes a wide range of structurally diverse compounds that can be divided into classes of primary (e.g., sterols, carotenoids, growth regulators, and the polyprenol substituents of dolichols, quinones, and proteins) and secondary (e.g., monoterpenes, sesquiterpenes, and diterpenes) metabolites. The primary metabolites are important for biological phenomena such as the preservation of membrane integrity, photoprotection, orchestration of developmental programs, and anchoring of essential biochemical activities to specific membrane systems; the secondary metabolites participate in processes involving inter-cellular communication, and appear to mediate interactions between plants and their environment (see, for example, Stevens, in *Isoprenoids in Plants* [Nes et al., eds], Macel Dekker et al., New York, pp. 65-80, 1984; Gibson et al., *Nature* 302:608, 1983; and Stoessl et al., *Phytochemistry* 15:855, 1976).

Isoprenoids are synthesized through the polymerization of isoprene building blocks, combined with cyclization (or other intramolecular bond formation) within intermediate or final product molecules. The polymerization reactions are catalyzed by prenyltransferases that direct the attack of an electron deficient carbon on the electron-rich carbon atom in the double bond on the isoprene unit (see Figure 14, from U.S. Patent Number 5,824,774). Cyclizations and other intramolecular bond formation reactions are catalyzed by isoprenoid, or terpene, synthases (see Figure 15, from U.S. Patent Number 5,824,774).

The terpene synthase proteins are modular proteins in which functional domains tend to correspond with natural exons (see, for example, U.S. Patent Number 5,824,774, incorporated herein by reference). Figure 16, from U.S. Patent Number 5,824,774, presents a schematic illustration of the correspondence between natural exons and functional domains within isoprenoid synthases. The upper diagram represents the organization of exons within the TEAS gene, which is nearly identical to that of the HVS and casbene synthase genes; the lower diagram shows the alignment of functional domains to the exonic organization of the TEAS and HVS genes.

As will be appreciated in light of the present application, the instant invention provides a system by which DNA molecules encoding isoprenoid synthase functional domains may be linked to one another to generate novel hybrid isoprenoid synthase genes in which the arrangement and/or number of functional domains is varied as compared with those observed in naturally-occurring isoprenoid synthase genes. These novel hybrid genes will encode novel hybrid proteins that are expected to catalyze the synthesis of new isoprenoid compounds.

As mentioned above, in some embodiments of the invention, DNA molecules encoding functional domains from one protein family are linked to DNA molecules encoding functional domains from a different protein family. Of particular interest in accordance with the present invention are reactions in which DNAs encoding polyketide synthase functional domains are linked with DNAs encoding peptide synthetase functional domains. Alternative preferred embodiments involve linkage of fatty acid synthase functional domains with either

or both of polyketide synthase functional domains and peptide synthetase functional domains. The hybrid genes created by such inter-family ligation reactions can then be tested according to known techniques to determine their ability to encode proteins that catalyze the synthesis of novel chemical compounds related to polyketides, fatty acids, and/or peptides.

As also mentioned above, it will be appreciated that the DNA molecules selected to be linked to one another in a particular experiment are not limited to molecules encoding functional domains or portions thereof; molecules encoding "linker" amino acids may additionally or alternatively be employed, as can non-coding molecules, depending on the desired final product.

To give but one example, it may sometimes be desirable to include in a final ligated molecule certain control sequences that will regulate expression of other DNA sequences to which the control sequences are linked when the ligated molecule is introduced into a host cell or an *in vitro* expression system. For example, transcriptional control sequences, RNA splicing control sequences, other RNA modification control sequences, and/or translational control sequences may be included in one or more of the DNA molecules to be linked together. A wide variety of such expression control sequences are well known in the art (see, for example, Sambrook et al., *Molecular Cloning: A Laboratory Manual*, 2nd Ed., Cold Spring Harbor Press, Cold Spring Harbor, New York, 1989, incorporated herein by reference); those of ordinary skill in the art will be familiar with considerations relevant to selecting desirable control sequences for use in their particular application. In general, so long as such control sequences direct expression of other DNA sequences to which they are linked when those DNA sequences are introduced into a cell or an *in vitro* expression system, they are appropriate for use in accordance with the present invention.

Other DNA modules that could desirably be used in accordance with the present invention include, for example, modules encoding a detectable protein moiety (e.g., an enzyme moiety that catalyzes a detectable reaction such as a color change or induction of fluorescence or luminescence, or a moiety that interacts with a known monoclonal antibody, etc), modules encoding a moiety that allows ready purification of any polypeptide encoded by the ligated product DNA molecule (e.g.,

a GST domain, a copper chelate domain, etc.), or any other module desired by the researcher.

Directional ligation

5 As discussed herein, one particularly valuable application of the inventive techniques is for the linkage of multiple different nucleic acid molecules to one another. Because the embodiments of the invention that provide product molecules with 3' or 5' overhangs allow the sequence and length of those overhangs to be selected at the practitioner's discretion, molecules can readily be prepared for
10 ligation only to certain designated partners, in certain designated orders, so that multi-member ligation reactions can be performed with only minimal generation of spurious or undesired ligation products.

 Figure 17 presents a schematic depiction of one generic example of such a directional ligation reaction according to the present invention (Figures 18-22 and
15 Example 1 describe a specific example). As shown, a first nucleic acid molecule, designated "A", contains a first overhang, designated "overhang 1" on one end. A second nucleic acid molecule, "B" is flanked by a second overhang, "overhang 1'", that is complementary to overhang 1, and a third overhang, "overhang 2", that is preferably unrelated to, and certainly not identical with, overhang 1. A third
20 nucleic acid molecule, "C", contains a fourth overhang, "overhang 2'", that is complementary with overhang 2. As will be appreciated by those of ordinary skill in the art, a ligation reaction including all three of these nucleic acid molecules will produce only a single reaction product, "ABC", and will not produce "AC" or circular "B" products due to the incompatibility of the ends that would have to be
25 ligated together to generate such products.

Mutagenesis

 In another particularly useful application, the inventive techniques and reagents may be utilized to alter the nucleotide sequence of nucleic acid molecules
30 that are being linked together. A separate mutagenesis reaction is not required. Rather, primers and/or overhangs whose sequence and length may be selected by the practitioner are utilized to create single-stranded regions between molecules to

be ligated, which single-stranded regions include new or altered sequences as desired. These single-stranded regions can subsequently be filled in with a polymerase that will synthesize a strand complementary to the new sequence. Alternatively or additionally, primers may be employed that add sequence to a particular product molecule strand that will be copied in an extension or amplification reaction.

Exon shuffling

One particular application of the techniques and reagents described herein is in the production of libraries of hybrid nucleic acid molecules in which particular collections of DNA molecules, or "exons" have been linked to one another. That is, an "exon shuffling" reaction is one in which a single reaction mixture (e.g., a ligation mixture or a splicing reaction-- discussed further below) generates at least two, and preferably at least 10, 100, 1000, 10,000, 100,000, or 1,000,000 different product molecules.

As used herein, the term "exon" refers to any DNA molecule that is to be ligated to another DNA molecule. An exon may include protein-coding sequence, may be exclusively protein-coding, or may not include protein-coding sequence at all. The term "exon shuffling" is intended to indicate that, using the techniques and reagents of the present invention, collections of exons can be produced that can be ligated to one another in more than one possible arrangement. For example, as depicted in Figure 23, the inventive techniques and reagents may be employed in a ligation reaction in which a single upstream exon, A, can be ligated to any one of a collection of different internal exons (B1-B4 in Figure 23), which in turn is further ligated to a downstream exon, C.

Those of ordinary skill in the art will readily appreciate that Figure 23 presents just one particular embodiment of an "identity exon shuffling" reaction (i.e., one in which the identity of a particular exon is different in different products of the shuffling reaction) according to the present invention. A wide array of related reactions is included within the inventive "exon shuffling" concept, and particularly within the concept of "identity exon shuffling". For example, more than one exon may be varied in a particular shuffling reaction. In fact, it is not

necessary to have upstream and downstream terminal exons that are uniform among shuffling products, as is depicted in Figure 23. Such consistency may provide certain advantages, however, including an ability to amplify all shuffling products with a single set of amplification primers (discussed in more detail below). Even if
5 invariant flanking exons are preserved, however, more than one internal exon may be varied; even if additional invariant internal exons are also provided.

Figure 24 presents an embodiment of a different sort of exon shuffling reaction that may be performed according to the present invention. In the particular embodiment shown in Figure 24, upstream (A) and downstream (H)
10 exons are provided in combination with a wide variety of possible internal exons (B-G). All exons have compatible overhangs. In such a reaction, the possibilities for internal exons arrangements to be found in product molecules are infinite. Also, because no exons (other than the optional flanking exons) are restricted to a particular position in the exon chain, this type of shuffling is referred to as
15 "positional shuffling".

Of course, those of ordinary skill in the art will appreciate that Figure 24 is but an exemplary embodiment of inventive positional shuffling systems. For example, it may well be desirable to employ at least two sets of compatible overhangs and to ensure that potential internal exons are not flanked by compatible
20 ends; otherwise, intramolecular circularization can present serious complications as a competing reaction in inventive ligations. Also, it is possible to perform an exon shuffling reaction that represents a compromise between the extremes of allowing identity shuffling at a single position while holding all other positions fixed (e.g., Figure 17) and allowing complete shuffling at all positions. Merely by selecting
25 the compatibility of the overhangs, the practitioner may limit the number of exons able to incorporate at a particular chain site, while allowing more variability at a different site.

One of the advantages of the present invention is that it allows simultaneous multi-site variation, optionally in combination with positional variation (i.e., the
30 possibility that a particular exon sequence could end up in different positions in different product molecules. To give but one example of the significance of this phenomenon, Figure 25 shows that other techniques might allow production of

libraries in which a single position in an exon chain can be varied at one time. For a three-exon chain in which 10 different exons could be employed at each of the positions, 30 different variants can be produced (A1BC, A2BC, A3BC, . . . A10BC, AB1C, AB2C, . . . AB10C, ABC1, ABC2, . . . ABC10). By contrast, if
5 all three positions can be varied simultaneously, as is possible in accordance with the present invention, 1000 different variants can be produced.

As discussed above, it is now accepted that the evolutionary process often produces new genes by re-sorting existing exons. Large gene families have apparently been produced by exon shuffling. According to the present invention, it
10 is desirable to employ the inventive techniques both to link particular selected functional domains to one another (see above) and to shuffle exons found in those gene families, so that a library of (at least two) product genes is generated.

The inventive exon shuffling techniques may be applied to any desired collection of exons. Preferably, they are applied to exons including protein-coding
15 sequences. More preferably, they are applied to protein-coding exons that have been re-used in evolution in different members of gene families (see discussion above). In one particularly preferred embodiment of the exon shuffling system of the present invention, the exons to be shuffled represent functional domains of synthetic enzymes. As discussed above with respect to ligation, re-sort exons from
20 within family or between or among families.

Particularly preferred gene families to which inventive exon shuffling techniques may be applied include, but are not limited to, the tissue plasminogen activator gene family, the animal fatty acid synthase gene family, the polyketide synthase gene family, the peptide synthetase gene family, and the terpene synthase
25 gene family. The class I bacterial polyketide synthase gene family presents a particularly attractive target for application of the inventive exon shuffling techniques in that the co-linearity of functional domains and catalytic capabilities is so well established for this family.

Also, the close mechanistic relationship between class I polyketide synthases and animal fatty acid synthases, class II polyketide synthases, and/or intermediate
30 class polyketide synthases (e.g., fungal polyketide synthases, whose functional organization and catalytic characteristics are apparently intermediate between those

of the bacterial class I and class II polyketide synthases) renders shuffling reactions that admix DNAs encoding functional domains of two or more of these different families particularly intriguing. Such reactions will generate libraries of new synthetic enzymes, which in turn will generate libraries of new chemical compounds that can be assayed according to any available technique to determine whether they have interesting or desirable biological activities.

Integration with existing technologies

It will be appreciated that the present invention does not describe the only available method for linking selected nucleic acid molecules to one another. For example, the established restriction-enzyme-based technology clearly allows cleavage and ligation of nucleic acid molecules, albeit without the convenience and other advantages of the inventive system. Also, techniques have been developed by which ribozymes can be employed to mediate cleavage and ligation of nucleic acids at the RNA or DNA level (see, for example, U.S. Patent Number 5,498,531; U.S. Patent Number 5,780,272; WO 9507351; WO 9840519, and U.S. Patent Application Serial Number 60/101,328, filed September 21, 1998, each of which is incorporated herein by reference; see also Example 4).

Each of these different systems for nucleic acid manipulation offers certain advantages and disadvantages. For example, ribozyme-mediated systems offer the distinct advantage that shuffling reactions may be performed *in vivo* if desired (see, for example, U.S. Patent Application Serial Number 60/101,328, filed September 21, 1998). Furthermore, once a shuffling cassette is generated in which an exon of interest is linked to a first trans-splicing ribozyme component, that exon may be ligated to any other exon that is linked to a second trans-splicing component that is compatible with the first trans-splicing component in a simple trans-splicing reaction. Thus, the more the ribozyme-mediated system is utilized, and the larger the number of shuffling cassettes generated by its use, the more powerful it becomes.

Ribozyme-mediated nucleic acid manipulation, like the techniques described herein, can be used for exon shuffling, and can be engineered to direct seamless ligation of any selected nucleic acid molecules. Furthermore, like the inventive

system, the ribozyme-mediated system may be engineered so that the agents that mediate ligation (the ribozyme components in the ribozyme-mediated system; the overhangs in the inventive system) are only compatible with certain selected other ligation-mediating agents. This ability allows one to perform directed ligation reactions analogous to those depicted in Figure 17, in which a collection of exons is incubated together but only certain selected exons can become ligated to one another (see, for example, Example 4 and Figure 29).

One particularly preferred embodiment of the present invention represents an integration of the primer-based manipulation techniques described herein with the ribozyme-mediated techniques described in the above-referenced patents and patent applications. Specifically, the primer-based nucleic acid manipulation techniques described herein are utilized to construct ribozyme-associated shuffling cassettes that are then employed in splicing reactions to generate hybrid nucleic acid molecules that can subsequently be cloned and manipulated using inventive primer-based strategies.

Figure 26 presents one version of such a combined primer-based/ribozyme-mediated nucleic acid manipulation scheme. As depicted, nine different product molecules are produced using inventive primer-based nucleic acid manipulation strategies. These molecules are designed to be ligated together to produce three different shuffling cassettes. The first shuffling cassette comprises (i) a promoter that will direct transcription of the cassette; (ii) a first tag sequence; (iii) an upstream terminal exon; and (iv) a first ribozyme component. The second shuffling cassette comprises (i) a promoter that will direct transcription of the cassette; (ii) a second ribozyme component, compatible with the first ribozyme component; (iii) an internal exon; and (iv) a third ribozyme component (optionally not compatible with the second ribozyme component). The third shuffling cassette comprises (i) a promoter that will direct transcription of the cassette; (ii) a fourth ribozyme component that is compatible with the third ribozyme component (and optionally not with the first ribozyme component); (iii) a downstream terminal exon; and (iv) a second tag sequence.

Given the ease with which shuffling cassettes may be generated using the inventive primer-based technology, there is no need for shuffling cassettes to be

introduced into vectors; they may be transcribed directly. Of course, they may be introduced into vectors if so desired, preferably by means of the inventive primer-based nucleic acid manipulation techniques. Each cassette is transcribed and the transcription products are incubated with one another under splicing conditions, either *in vitro* or *in vivo*, to produce a hybrid molecule containing each of the three exons. The hybrid molecule may then be introduced into a vector or further manipulated, again preferably using the inventive primer-based manipulation technology.

Those of ordinary skill in the art will appreciate that more than one internal cassette may be employed in the system of Figure 26, either in an exon shuffling (involving positional and/or identity shuffling) reaction or in a directed ligation reaction in which only one copy of each exon will be introduced into the hybrid molecule, in a pre-determined order. Alternatively or additionally, multiple alternative upstream or downstream exons may be employed, or such terminal exons may be left out. In a particularly preferred embodiment of an identity exon shuffling reaction, multiple alternative exons are provided, and are simultaneously shuffled, for at least two positions (e.g., one internal position and one terminal position, two internal positions, or two terminal positions) in the hybrid molecule.

One advantage of the combined primer-based/ribozyme-mediated system depicted in Figure 26 can be appreciated through consideration of the number of primers required to generate the indicated molecules, and/or to clone them into vectors or other desirable locales, according to the inventive methods. For example, sixty-seven primers are required to generate the initial product molecules if 10 different possible exon product molecules are produced for each of the "A", "B", and "C" exons. This is a relatively large number of primers, but is justified by the ease with which the product molecules are generated and ligated together using the inventive system, as compared with alternative methods (e.g., standard restriction-enzyme-based cloning techniques) available for the production of the shuffling cassettes. Only four primers are required to amplify the resulting shuffling cassettes, or to ligate them to other DNA molecules (e.g., a vector). Most importantly, only two primers are required to amplify (or ligate) assembled genes. Particularly where exon shuffling reactions have been performed, and a library of

assembled genes is generated, it is valuable to be able to amplify all members of the library with the same two primers.

Automation

5 One particularly attractive feature of the inventive techniques and reagents is their susceptibility to automation. In particular, where large libraries of novel hybrid nucleic acids are being produced in inventive exon shuffling reactions, it may be desirable to employ an automated system (e.g., the Beckman 2000 Laboratory Automation Work Station) to accomplish the simultaneous manipulation
10 of a large number of different samples.

 To give but one example of a preferred automated application of the present inventive methods, Figure 27 depicts a robotic system that could be utilized, for example, to accomplish exon shuffling as depicted in Figure 27 and further to screen the products of the shuffling reaction for desired activities. For example,
15 the product molecules depicted in the first column of Figure 26 could be generated by PCR in 96 well plates using a Biomek 2000 system in combination with a multimek 96 automated 96-channel pipetter and a PTC-225 DNA engine (MJ Research), relying on the ORCA robot arm to move the plates from one location to another as necessary.

20 Preferably, multiple alternatives are simultaneously prepared of each exon product molecule (e.g., n "A" exons, A1-An, are prepared; as are x "B" exons, B1-Bx; and y "C" exons, C1-Cy), along with T7/X, 1-4', T7/5,6, and Y products. As discussed above, 67 different primers are required to produce these product molecules according to the inventive methodologies described herein.

25 The automated system is then programmed to pipette the appropriate product molecules together, along with desired ligation reagents, to produce 30 shuffling cassettes of the types depicted in the second column of Figure 26. The system is then programmed to generate RNA from these shuffling cassettes using T7 RNA polymerase. The "A"-type, "B"-type, and "C"-type transcripts are then
30 mixed together in all possible combinations, and are incubated (still in the robotic system) under trans-splicing conditions. All together, 1000 different splicing reactions will be performed.

A small aliquot of each splicing reaction is then removed and amplified with inventive primers so that the amplification products can readily be ligated with a recipient molecule such as a vector. The resulting plasmids may then be introduced into host cells (e.g., bacterial cells) for further amplification, or
5 alternatively may be introduced into an *in vivo* or *in vitro* expression system so that any protein products encoded by the assembled shuffled genes may be assayed. Desirable expression systems will depend on the nature of the nucleic acid sequences that were shuffled. To give but one example, if fungal polyketide synthase gene fragments (e.g., encoding functional domains of fungal polyketide
10 synthase proteins) were shuffled according to this approach, it may be desirable to express the hybrid proteins thereby generated in one or more fungal or mammalian cells types in order to assess their synthetic capabilities.

Kits

15 Reagents useful for the practice of the present invention may desirably be provided together, assembled in a kit. Certain preferred kits will include reagents useful for both primer-mediated and ribozyme-mediated nucleic acid manipulation reactions.

Examples

Example 1

Preparation and Ligation of Product Molecules with 5' Overhang Sequences

This Example describes the preparation and ligation of product molecules having 5' overhangs, using hybrid primers containing deoxyribonucleotides at their
25 3' ends and ribonucleotides at their 5' ends.

Figure 18 presents a schematic of the particular experiment that was performed. As shown, three different product molecules were generated, two of which correspond to exons of the gene for subunit B of the human glutamate receptor, and one of which corresponds to an intron from the unrelated human β -
30 globin gene. The particular glutamate receptor exons we utilized are known as Flip and Flop, and are indicated in Figure 19A and 19B, which present the nucleotide

sequences of each of these exons (GenBank accession numbers X64829 and X64830, respectively).

We prepared each of our three product molecules by PCR, using Vent® DNA polymerase and plasmids Human GluR-B #7 (a cloned genomic fragment containing exons 13-16 of the human glutamate receptor B subunit) or HβT7 (a cloned genomic fragment containing exons 1-2 of the human β-globin gene).

The Flop exon was amplified with a 5' primer (primer 1 in Figure 18; 5'-AAATGCGGTAACTCGCAG, SEQ ID NO:____) that is entirely DNA and corresponds to the first 20 bases of the Flop exon, in combination with a 3' primer (primer 2 in Figure 18; 5'-accuTGAATCACCTCCCCC SEQ ID NO:____) whose 5'-most four residues are RNA, as indicated by lower case letters in Figure 18. This primer corresponds to the last 18 bases of the Flop exon plus 2 bases of intron. Together, these primers amplify a fragment corresponding to all of the human glutamate receptor Flop exon (115 basepairs) plus the first two residues at the 5' end of the intron.

Intron 1 was amplified with a 5' primer (primer 3 in Figure 18; 5'-agguTGGTATCAAGGTTACA, SEQ ID NO:____) whose sequence corresponds to the first 18 bases of the human β-globin intron 1, and whose 5'-most four residues are RNA, and are complementary to the four RNA residues at the 5' end of primer 2; in combination with a 3' primer (primer 4 in Figure 18, 5'-cuAAGGGTGGGAAAATAGAC, SEQ ID NO:____) corresponding to the last 20 bases of the human β-globin intron 1, whose 5'-most two residues are RNA. These primers together amplify a fragment corresponding to the entire intron (129 bp), and 2 add two residues corresponding to the last two residues at the 3' end of the Flop exon.

The Flip exon was amplified with a 5' primer (primer 5 in Figure 18, 5'-agAACCCCAGTAAATCTTGC, SEQ ID NO:____) corresponding to the first 18 bases of the human glutamate receptor Flip exon, whose 5'-most two residues are RNA and are complementary to the two RNA residues at the 5'-end of primer 4; in combination with a 3' primer (primer 6 in Figure 18, 5'-CTTACTTCCCGAGTCCTTGG, SEQ ID NO:____) corresponding to the last 20 exon bases, that was entirely DNA. These primers together amplify a fragment

corresponding to the entire Flip exon (115 bp) and the last two nucleotides at the 3' end of the intron.

Each amplification reaction included 400 μ mole of each primer, kinased (using T4 polynucleotide kinase in 100 μ l 1 X NEB T4 ligase buffer [50 mM Tris-HCl pH 7.8, 10 mM MgCl_2 , 10 mM DTT, 1 mM ATP, 25 μ g/ml BSA] for 30 minutes at 37 $^\circ\text{C}$, followed by dilution to 10 pmol/ μ l with 200 μ l nuclease-free dH_2O); 2 units Vent_R° (exo-) polymerase (NEB, Beverly, MA), 100 μ l 1 X Vent buffer (10 mM KCl, 10 mM $(\text{NH}_4)_2\text{SO}_4$, 20 mM Tris, 2 mM MgSO_4 , 0.1% Triton X-100); 200 μ M dNTPs; and 5 ng of template plasmid. One cycle of (i) 95 $^\circ\text{C}$, 3 minutes; (ii) 60 $^\circ\text{C}$, 3 minutes; (iii) 72 $^\circ\text{C}$, 3 minutes was followed by 35 cycles of (i) 95 $^\circ\text{C}$, 15 seconds; (ii) 60 $^\circ\text{C}$, 15 seconds; (iii) 72 $^\circ\text{C}$, 30 seconds, in a Robocycler[®] gradient 40 (Stratagene, La Jolla, CA) thermocycler.

We found that Vent_R° and Vent_R° (exo-) did not copy the ribonucleotides in our primers, so that, after amplification, each product molecule contained a 5' ribonucleotide overhang at one or both ends (4 nucleotides at the 3'-end of the Flop product; 4 nucleotides at the 5'-end of the β -globin intron product; 2 nucleotides at the 3'-end of the β -globin intron product; and 2 nucleotides at the 5'-end of the Flip product).

Each amplified product was precipitated with ethanol (EtOH) and was resuspended in 10 μ L, 2 of which were run on a 6% polyacrylamide gel in order to verify the presence of all three amplification products. Aliquots (2-4 μ L each) containing approximately equimolar quantities of each fragment were then combined in a ligation reaction containing 1 X New England Biolabs (NEB) T4 ligase buffer (50 mM Tris, pH 7.8, 10 mM MgCl_2 , 10 mM DTT, 1 mM ATP, 25 μ g/ml BSA) and 0.5 U of T4 DNA ligase (NEB, Beverly, MA). The 20 μ L reaction was incubated overnight at 4 $^\circ\text{C}$ to allow ligation to occur. Products of ligation were then amplified using primers 1 and 3 and *Taq* polymerase, which does copy RNA (Myers et al., *Biochem.* 6:7661, 1991). The amplification reaction contained 1 X *Taq* buffer (20 mM Tris, pH 9.0, 50 mM KCl, 0.1% Triton X-100), 200 μ M dNTPs, 5 Units of *Taq* polymerase (Promega, Madison, WI), 2 μ L of the ligation mix, and 400 μ mol of each primer.

The product of the *Taq* amplification is shown in Figure 20, and was ligated into the PCR 2.1 vector (Invitrogen, Carlsbad, CA) using the TA Cloning Kit according to manufacturer's instructions. Sequence analysis (using standard dideoxy sequencing methods, and Universal and Reverse primers from United States Biochemical, Cleveland, Ohio) of multiple (9) clones confirmed that all ligation junctions were correct (see Figures 21 and 22). Because this strategy ligated product molecules with rubonucleotide overhangs, it is sometimes referred to as Ribonucleotide overhang cloning (ROC).

Example 2

Preparation and Ligation of Product Molecules with 3' Overhang Sequences

This Example describes the preparation and ligation of product molecules having 3' overhangs, using hybrid primers containing deoxyribonucleotides at their 3' ends and ribonucleotides at their 5' ends.

Figure 28 presents a schematic of the particular experiment that was performed. As shown, three different product molecules were generated, two of which correspond to the Flip and Flop exons of the gene for subunit B of the human glutamate receptor, and one of which corresponds to an intron from the unrelated human β -globin gene (see Example 1).

Each of the three product molecules was prepared by PCR, using a *Pfu* polymerase which copies RNA nucleotides, and either human genomic DNA or HBT7 (see Example 1). The Flop exon was amplified with primers 1 and 2 from Example 1; intron 1 was amplified either with primers 3 and 4 from Example 1 or with primer 3 and an alternative primer 4 (5'uucuAAGGGTGGGAAAATAG-3'; SEQ ID NO: _____); the Flip exon was amplified either with primers 5 and 6 or with an alternative primer 5 (5'agaaCCCAGTAAATCTTGC; SEQ ID NO: _____); and primer 6.

Each 100 μ L reaction contained 2.5 U of *Pfu* Turbo[®] polymerase (Stratagene), 1X Cloned *Pfu* buffer (10 mM (NH₄)₂SO₄, 20 mM Tris pH = 8.8, 2 mM Mg SO₄, 10 mM KCl, 0.1 % Triton X-100 and 0.1 mg/ml BSA), 200 μ M of each dNTP, 1 mM MgSO₄, and primers at a final concentration of 0.5 μ M each. The Flop and Flip reactions contained 375 ng of human genomic DNA, while the

β-globin reaction contained 5 ng of HBT7 DNA. The PCR step program was one cycle of 95 °C, 5 min; 50 °C, 3 min; 72 °C 3 min; followed by 40 cycles of 95 °C, 30 sec; 50 °C, 30 sec; 72 °C, 45 sec; followed by one cycle of 72 °C, 5 min in Robocycler gradient 40 for the Flip and Flop fragments. The same program was used to amplify β-globin intron 1, except the annealing temperature was 46 °C. Since *Pfu* polymerase does not copy RNA (stratagene product literature), the PCR product literature), the PCR products contained 5' overhangs. These overhangs were filled in during an incubation at 72 °C for 30 minutes with 5 U of *Tth* polymerase (Epicentre Technologies, Madison, WI), to fill in the 5'-RNA overhangs (Note, in more recent experiments, M-MLV RT was used, rather than *Tth*, to fill in the overhangs. When M-MLV RT was used, the fragments were separated on agarose gels prior to treatment with 200 U of M-MLV RT in 1X First strand buffer (50 mM Tris pH = 8.3, 75 mM KCl, 3 mM MgCl₂), 10 mM DTT and 0.5mM dNTP in 20 μL.). This strategy allowed us to use *Pfu* polymerase, which has the highest fidelity of available thermostable DNA polymerases, during the amplification reaction but still generate blunt-ended reaction products.

The amplified parental PCR products were excised from an agarose gel and purified. Five μl of each purified sample were fractionated on an agarose gel for quantitation. We then converted the blunt-ended products to products containing 3' overhangs by removing the ribonucleotides through exposure to mild base. NaOH (1 N) was added to 8 μl of each of the gel isolated fragments to a final concentration of 0.2 N and the samples were incubated at 45 °C for 30 min. The base was neutralized by addition of 2 μl of 1 N HCl. Since NaOH hydrolysis generates a 3'-phosphate and a 5'-OH, we had to phosphorylate the products to be able to ligate them. The DNA fragments were phosphorylated in 1X T4 ligase buffer (USB) in a total of 20 μl for 30 min at 37 °C using 10 U of PNK (USB). Approximately 25 ng (3-6 μl) of each phosphorylated product were combined in a final volume of 20 μl and ligated for 16 hours at 14 μC in 1X T4 ligase buffer with 5 Weiss U of T4 DNA ligase (USB).

To produce the chimeric Flop-β-Flip product, a secondary PCR amplification was performed as described above for the primary PCRs using 1 μl of ligation reaction as template, primers 1 and 6, and an annealing temperature of 58 °C. A

chimeric product of the expected size (360 bp) was observed. This product was cloned and sequenced; both ligation junctions were correct in 6 of 8 clones that were sequenced. Two clones each had an error at one of the ligation sites. In one clone, three base pairs were lost at the boundary between the β -globin intron and Flip. In the other clone, an A was changed to a T (data not shown). We suspect that *Tth* polymerase introduced these errors during the fill in step of the procedure. Because the strategy described here involved ligation of molecules containing DNA overhangs, it is sometimes referred to as DNA Overhang Cloning (DOC).

Example 3

Bioassays for Determining Success of Primer Copying and/or Ligation

The present Example describes techniques that could be used to evaluate the ability of a particular DNA polymerase to copy (i.e., to use as a template) a particular modified oligonucleotide primer. For example, the techniques described herein might be useful to determine whether a particular modified nucleotide or ribonucleotide (or collection thereof) can be replicated by one or more DNA polymerases.

Figure 29 presents one embodiment of the present bioassay techniques. As shown, two primers are provided that hybridize with a template molecule. The first primer is known to be extendible by a particular DNA polymerase; the second primer includes one or more modified nucleotides or ribonucleotides whose ability to block replication by the DNA polymerase is unknown. Any nucleotide modification may be studied in the system.

As shown in Figure 29, both primers are extended, so that, if replication is blocked, a product molecule with a 5' overhang is produced; a blunt-ended product molecule (or a molecule containing a single-nucleotide 3'-overhang, depending on the DNA polymerase employed) is generated if replication is not blocked.

The product molecule is then incubated with a vector containing a complementary 5' overhang and carrying a selectable marker (or a marker identifiable by screening). Only if replication was blocked will hybridization occur. Ligation is then attempted and should succeed unless the particular modification interferes with ligation of a nick on the complementary strand

(unlikely) or the modification is present at the 5' end of the overhang and is of a character that interferes with ligation to an adjacent 3' end. In order to simplify the experiment and minimize the number of variables in any particular reaction, it is expected that modifications will only be incorporated at the very 5' end of a primer if their ability to block replication is already known and the desire is to assess only their ability to interfere with ligation.

The ligation product is then introduced into host cells, preferably bacteria. Selectable (or otherwise identifiable) cells will grow and proliferate only if the modification in question did block replication and either (i) did not block ligation on the complementary strand; or (ii) did block ligation on the complementary strand but did not block *in vivo* nick repair. If the modification were at the 5' end of the primer, cells will only grow if the modification did block replication and did not block ligation of both strands.

Of course, where the modification constitutes one or more ribonucleotides, or other removable nucleotides, absence of colonies due to inability to block replication can be distinguished from other absence of colony results by treating the original product molecule with an agent that will remove the modified nucleotide(s), along with any more 5' nucleotides, and then incubating the resulting secondary product molecule, which contains a 3' overhang complementary to the modified nucleotide and any more 5' nucleotides, with a vector containing a compatible 3' overhang.

Example 4

Directional Ligation of Multiple Nucleic Acid Molecules by Engineered Selective Compatibility of Catalytic Ribozyme Elements

Figure 30 shows a directional ligation reaction that allowed selective ligation of particular exons through use of incompatible ribozyme components. As indicated, transcripts were generated in which (i) a first exon (A) was linked to a first ribozyme component from the $\alpha I5\gamma$ group II intron; (ii) a second exon (B) was flanked by (a) a second ribozyme component, also from the $\alpha I5\gamma$ group II intron, that is compatible with the first ribozyme component, and (b) a third ribozyme component, from the LTRB intron of *Lactococcus lacti*, that is not compatible with

the second intron component; and (iii) a third exon was linked to a fourth ribozyme component, also from the LTRB intron, that is compatible with the third intron component but not with the first intron component. These three transcripts were incubated together under splicing conditions and, as shown, only the ABC product (and not the AC nor the circular B product) was produced.

In all, nine plasmids were used in the study: pJD20, pB.E5.D4, pD4.E3(dC).B(2), pLE12, pB.5'Lac, p3'Lac.B, pD4.E3(dC)B(2).5'Lac, and p3'Lac.B.E5.D4. Two PCR amplifications were performed using plasmid pJD20, which contains the full-length $\alpha 5\gamma$ intron (Jarrell et al., *Mol. Cell. Biol.* 8:2361, 1988), as a template. The first reaction amplified part of the intron (domains 1-3 and 73 nt of domain 4), along with part (27 nt) of the 5' exon. The primers utilized, BamHI.E5 (5'-ACGGGATCCATACTTACTACGTGGTGGGAC; SEQ ID NO:____) and D4.SalI (5'-ACGGTCGACCCTCCTATCTTTTTTAATTTTTTTTTT; SEQ ID NO:____), were designed so that the PCR product had unique *Bam*HI and *Sal*I sites at its ends. The PCR product was digested with *Bam*HI and *Sal*I, and was ligated into the PBS- vector (Stratagene), digested with the same enzymes, so that it was positioned downstream of the T7 promoter. The resulting plasmid was designated pB.E5.D4, and encodes the B.5' γ shuffling cassette (see Figure 31).

The second PCR reaction that utilized pJD20 as a template amplified a different part of the intron (the remaining 65 nt of domain 4 plus domains 5-6), along with part (29 nt) of the 3' exon. The primers utilized, KpnI.D4 (5'-ACGGGTACCTTTATATATAACTGATAAATATTTATT; SEQ ID NO:____) and E3.BamHI (5'-ACGGGATCCAGAAAATAGCACCCATTGATAA; SEQ ID NO:____), were designed so that the PCR product had unique *Kpn*I and *Bam*HI sites at its ends. The PCR product was digested with *Kpn*I and *Bam*HI, and was ligated into the PBS- vector, digested with the same enzymes, so that it was positioned downstream of the T7 promoter. The resulting plasmid was called pD4.E3(dC).B (see Figure 31).

Sequence analysis of the pD4.E3(dC).B plasmid revealed an unexpected point mutation in the 3' exon sequence. The expected sequence was ACTATGTATTATCAATGGGTGCTATTTTCT (SEQ ID NO:____); the observed sequence was ACTATGTATTATAATGGGTGCTATTTTCT (SEQ ID NO:____).

A site directed mutagenesis reaction was then performed, using the QuickChange[®] Site-Directed Mutagenesis Kit (Stratagene, catalog number 200518) to insert an additional *Bam*HI site into the 3' exon sequence. The primers utilized were designated E3.BamHI(2) (5'-

5 CTCTAGAGGATCCAGAAAATAGGATCCATTATAATACATAGTATCCCG;
SEQ ID NO:____) and E3.BamHI(2)complement (5'-
CGGGATACTATGTATTATAATGGATCCTATTTTCTGGATCCTCTAGAG;
SEQ ID NO:____). The plasmid generated as a result of the site-directed
mutagenesis reaction was designated pD4.E3.(dC).B(2), and encoded the 3'γ.B
10 shuffling cassette (see Figure 31), in which the length of the 3' exon was shortened
to 13 nt.

Two additional PCR reactions were performed, in which the plasmid pLE12, which encodes the full-length LTRB intron flanked by its natural 5' and 3' exons (Mills et al., *J Bacteriol.* 178:3531, 1996), was used as a template. In the
15 first reaction, primers 5'transM.E.5' (5'-
CACGGGATCCGAACACATCCATAACGTGC; SEQ ID NO:____) and 5'sht3' (5'-
CAGCGTCGACGTACCCCTTTGCCATGT; SEQ ID NO:____) were used to
amplify part of the LTRB intron (domains 1-3), and part (15 nt) of the 5' exon.
The PCR product was generated with *Taq* polymerase and was cloned into the
20 PCR2.1 Topo vector (Invitrogen) using the Topo[®] TA Cloning[®] kit (Invitrogen).
The resulting plasmid was designated pB.5'Lac, and encodes the B.5'Lac shuffling
cassette (see Figure 31).

The same PCR product was also digested with *Bam*HI and *Sal*I, and was
ligated into pD4.E3(dC).B(2), cut with the same enzymes, to produce
25 pD4.E3(dC)B(2).5'Lac, which encodes the 3'γ.B.5'Lac shuffling cassette (see
Figure 31).

Additionally, plasmid pB.5'Lac was digested with *Spe*I and *Asp*718 to
remove some unwanted restriction sites. Overhangs were filled in with Klenow
fragment, and the resulting blunt ends were ligated to reseal the vector. The
30 plasmid thereby produced was designated pB.5'Lac(K) (see Figure 31).

The second PCR reaction that utilized pLE12 as a template involved the use
of primers 3'transM.E.5' (5'-

CACGGAGCTCTTATTGTGTACTAAAATTAAAAATTGATTAGGG; SEQ ID NO: _____) and 3'transM.E.3' (5'-

CAGCGGATCCCGTAGAATTAAAAATGATATGGTGAAGTAG; SEQ ID NO: _____) to amplify part of the PTRB intron (domains 4-6), attached to part (21 nt) of the 3' exon. The primers were designed so that the PCR product had unique *SacI* and *BamHI* sites at its ends. The PCR products was generated with *Taq* polymerase and was cloned into the pCR2.1 Topo vector. The resulting plasmid was designated 3'Lac.B, and encoded the 3'Lac.B shuffling cassette (see Figure 31).

Plasmid p3'Lac.B was digested with *SacI* and *BamHI*, and the 1993 bp band thereby generated was purified from an agarose gel using the Geneclean II kit (BIO 101). The purified fragment was then ligated into pE5.D4, digested with the same enzymes, to produce plasmid p3'Lac.B.E5.D4, encoding the 3'Lac.B.5'γ shuffling cassette (see Figure 31).

Plasmids pB.E5.D4, pD4.E3(dC).B(2), pB.5'Lac, p3'Lac.B, and pD4.E3(dC).B(2).5'Lac were linearized with *HindIII* and were transcribed *in vitro* with T7 RNA polymerase (Stratagene, catalog number 600123) at 40 °C for 1 hour in 100 μL reactions containing 6 μg of linearized template DNA and 0.5 mM unlabeled ATP, CTP, GTP, and UTP. The RNAs produced in these transcription reactions were treated with 1 U of RQ1 RNase-free DNase, were extracted with phenol-chloroform, were desalted on a Sephadex G25 column, and were precipitated with EtOH. Precipitates were subsequently resuspended in 6 μL water.

One μL of each resuspended RNA transcript was then used in a trans-splicing reaction carried out at 45 °C for 60 minutes, in 40 mM Tris-HCl, pH 7.6, 100 mM MgCl₂, and either 0.5 M NH₄Cl or 0.5M (NH₄)₂SO₄.

After the trans-splicing reaction, a reverse transcription/PCR reaction was performed to identify ligated splicing products. The detected products were: (i) ligated αγ5 exons E5 and E3 produced by trans-splicing of B.E5.D4 and D4.E3(dC).B(2) (lane 1, Figure 32); (ii) ligated LTRB 5' and 3' exons produced by trans-splicing of 3'Lac.B and 3'Lac.B (lane 2, Figure 33); and (iii) the three-molecule ligation product produced by trans-splicing of B.E5.D4, D4.E3(dC).B(2).5'Lac, and 3'Lac.B (lanes 2 and 3, Figure 33).

Example 5

Cloning Products of 3'-overhang Product Ligation without Amplification of Chimeric Product.

5 We found that the products of a DOC ligation reaction could be cloned directly into a vector for replication in bacteria without a chimeric amplification step. As was described above in Example 2, we designed chimeric primers that, when used in a DOC experiment, generated Flop, intron 1, and Flip PCR products that could be ligated directionally. In addition, the primers were designed such that NaOH treatment of the PCR products creates an upstream overhang on the Flop
10 exon that is compatible with an *Apa* I overhang, and a downstream overhang on the Flip exon that is compatible with a *Pst* I overhang. All three fragments were incubated together in the presence of ligase and pBluescript II SK (-) that had been digested with *Apa*I and *Pst*I. An aliquot of the ligation mixture was transformed directly into *E. coli*, and the expected chimeric clone was readily isolated,
15 sequenced, and found to be perfect (data not shown).

Example 7

Construction of Multiple Chimeric Products by DNA-Overhang Cloning

To demonstrate the generality of the procedures described herein, we
20 applied the techniques of Example 2 and to a variety of different molecules and produced five different chimeras, shown in Figure 35. All five chimeras were generated by directional three-molecule ligation. Note that these chimeras were generated using M-MLV reverse transcriptase, rather than *Tth*, to fill in 5' RNA overhangs. When M-MLV RT was used, no errors were detected at any of the
25 ligation points.

Other Embodiments

Those of ordinary skill in the art will appreciate that the foregoing has been a description merely of certain preferred embodiments of the present invention; this
30 description is not intended to limit the scope of the invention, which is defined with reference to the following claims:

We claim:

Claims

1. A double stranded DNA molecule with a single stranded overhang comprised of RNA.
- 5 2. A library of nucleic acid molecules, wherein each member of the library comprises:
at least one nucleic acid portion that is common to all members of the library; and
at least two nucleic acid portions that differ in different members of
10 the library.
3. The library of claim 2 wherein each of the nucleic acid portions comprises protein-coding sequence and each library member encodes a continuous polypeptide.
- 15 4. The library of claim 3 wherein each of the variable nucleic acid portions encodes a functional domain of a protein.
5. The library of claim 4 wherein the functional domain is one that is naturally
20 found in a gene family selected from the group consisting of the tissue plasminogen activator gene family, the animal fatty acid synthase gene family, the polyketide synthase gene family, the peptide synthetase gene family, and the terpene synthase gene family.
- 25 6. A method of generating a hybrid double-stranded DNA molecule, the method comprising steps of:
providing a first double-stranded DNA molecule, which double-stranded DNA molecule contains at least one single stranded overhang comprised of RNA;
30 providing a second double-stranded DNA molecule containing at least one single-strand overhang that is complementary to the RNA overhang on the first double-stranded DNA molecule; and

ligating the first and second double-stranded DNA molecules to one another so that a hybrid double-stranded DNA molecule is produced.

- 5 7. A method of generating a hybrid double-stranded DNA molecule, the method comprising:

generating a first double-stranded DNA molecule by extension of first and second primers, at least one of which includes at least one base that is not copied during the extension reaction so that the extension reaction produces a product molecule containing a first overhang;

- 10 providing a second double-stranded DNA molecule containing a second overhang complementary to the first overhang; and

ligating the first and second double-stranded DNA molecules to one another, so that a hybrid double-stranded DNA molecule is produced.

- 15 8. A method of generating a hybrid double-stranded DNA molecule, the method comprising:

generating a first double-stranded DNA molecule by extension of first and second primers, at least one of which includes at least one potential point of cleavage;

- 20 exposing the first double-stranded DNA molecule to conditions that result in cleavage of the cleavable primer at the potential point of cleavage, so that a first overhang is generated on the first DNA molecule;

providing a second double-stranded DNA molecule containing a second overhang complementary to the first overhang; and

- 25 ligating the first and second double-stranded DNA molecules to one another, so that a hybrid double-stranded DNA molecule is produced.

1/37

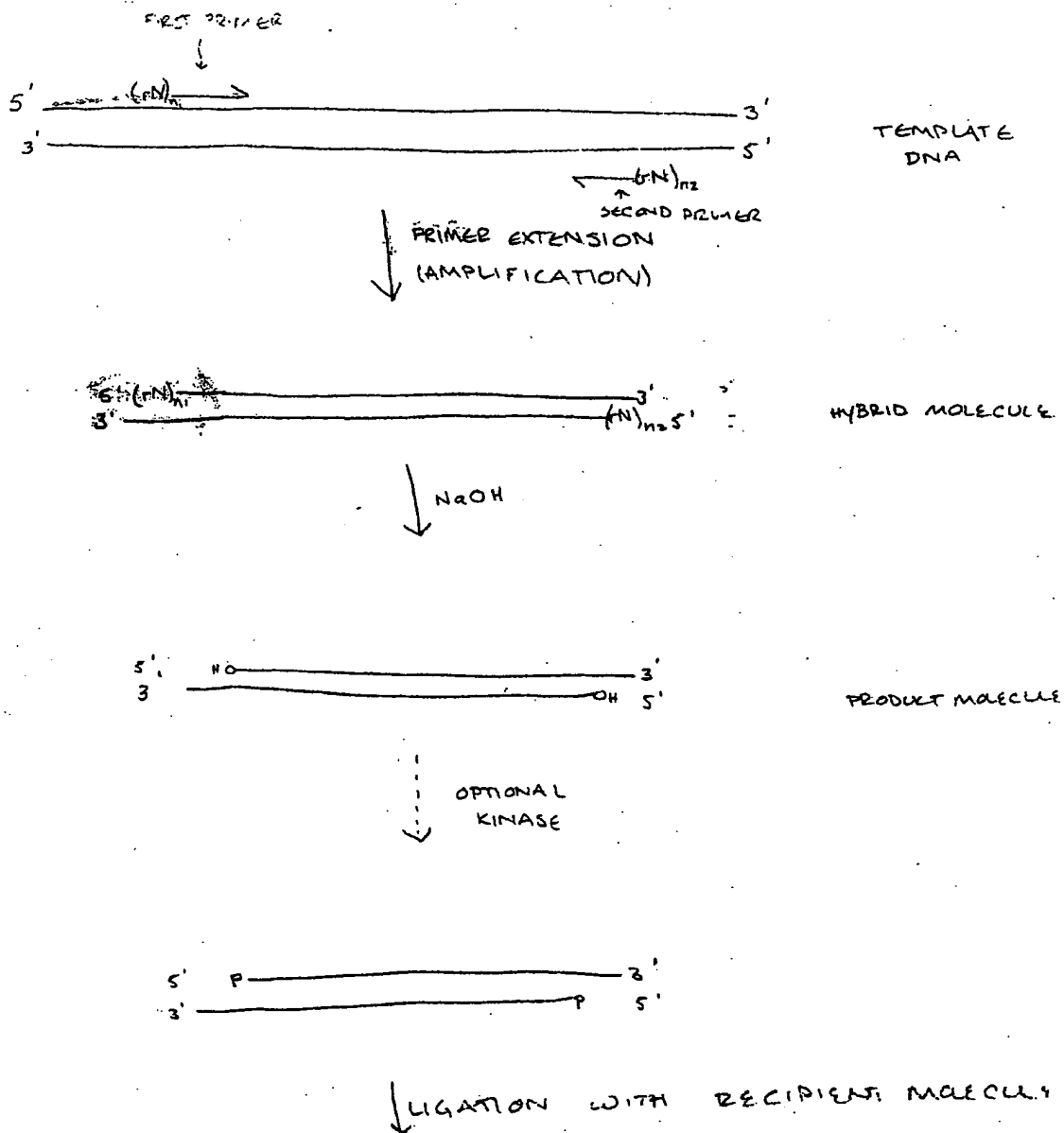


FIGURE 1

2/37

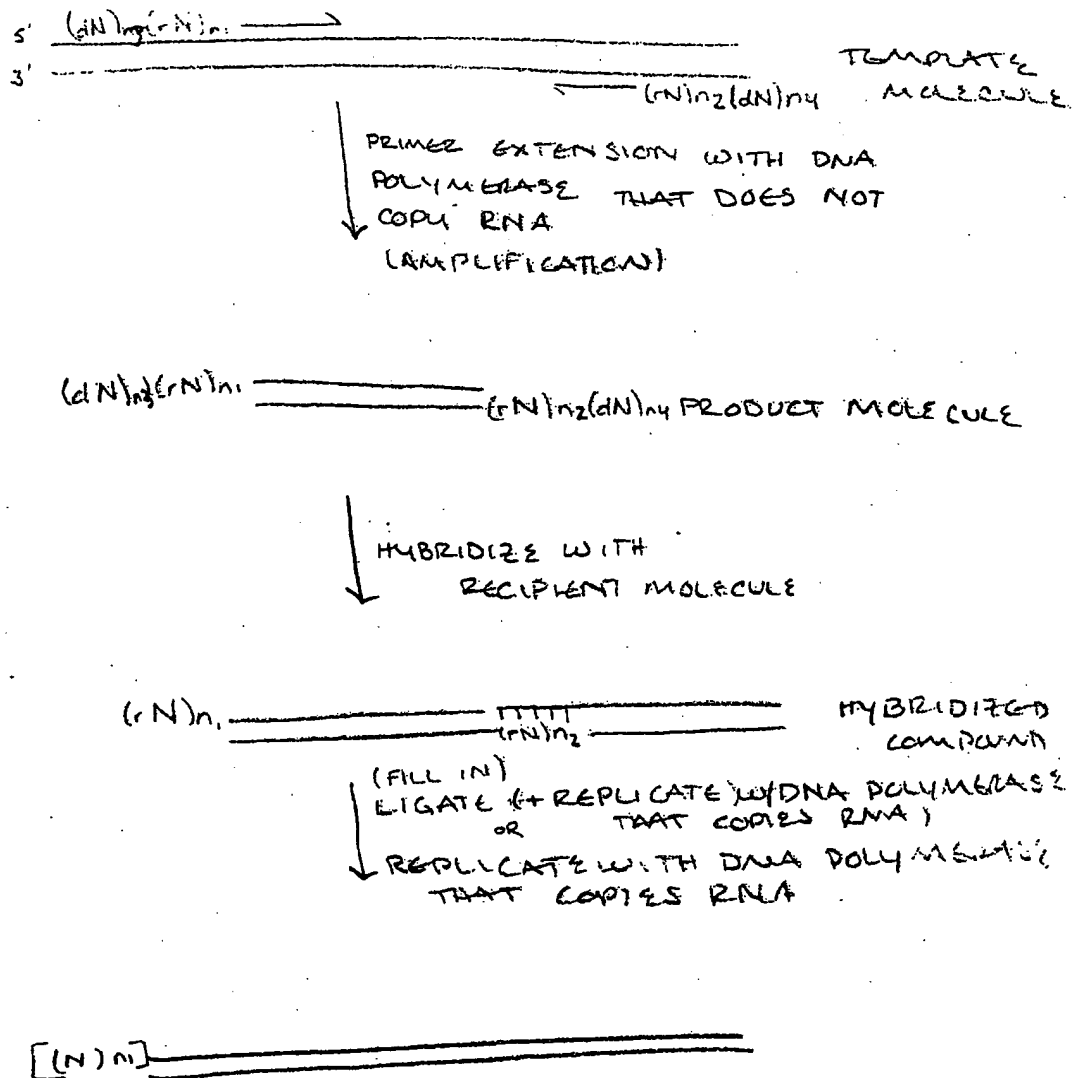
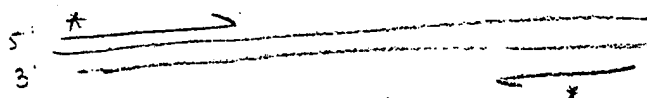


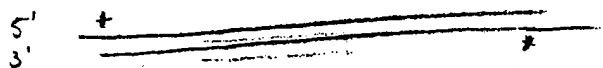
FIGURE 2

3/37



TEMPLATE
MOLECULE

↓
EXTENSION
(AMPLIFICATION)



PRODUCT MOLECULE

FIGURE 3

4/37

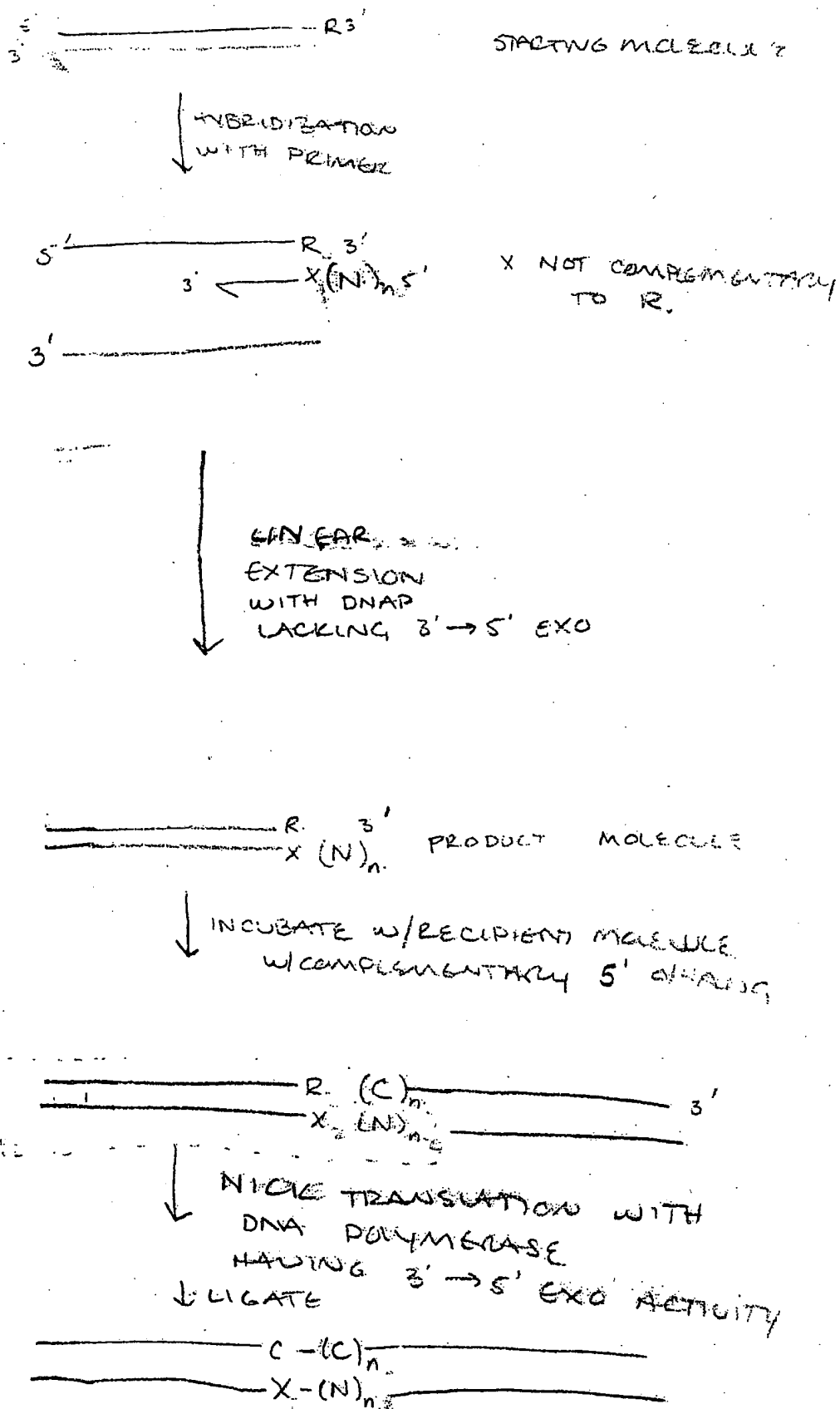


FIGURE 4A

5/37

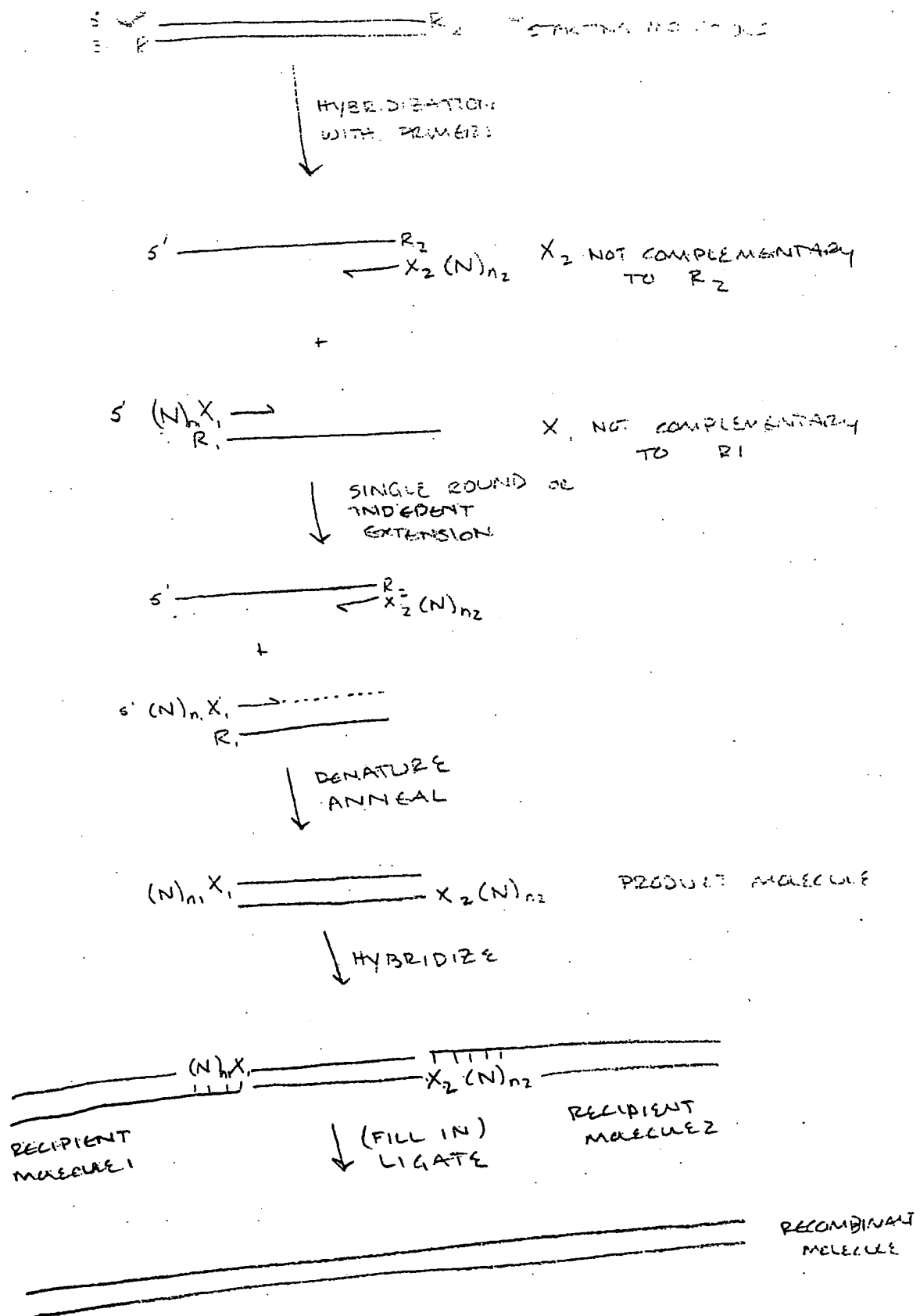


FIGURE 4B

6/37

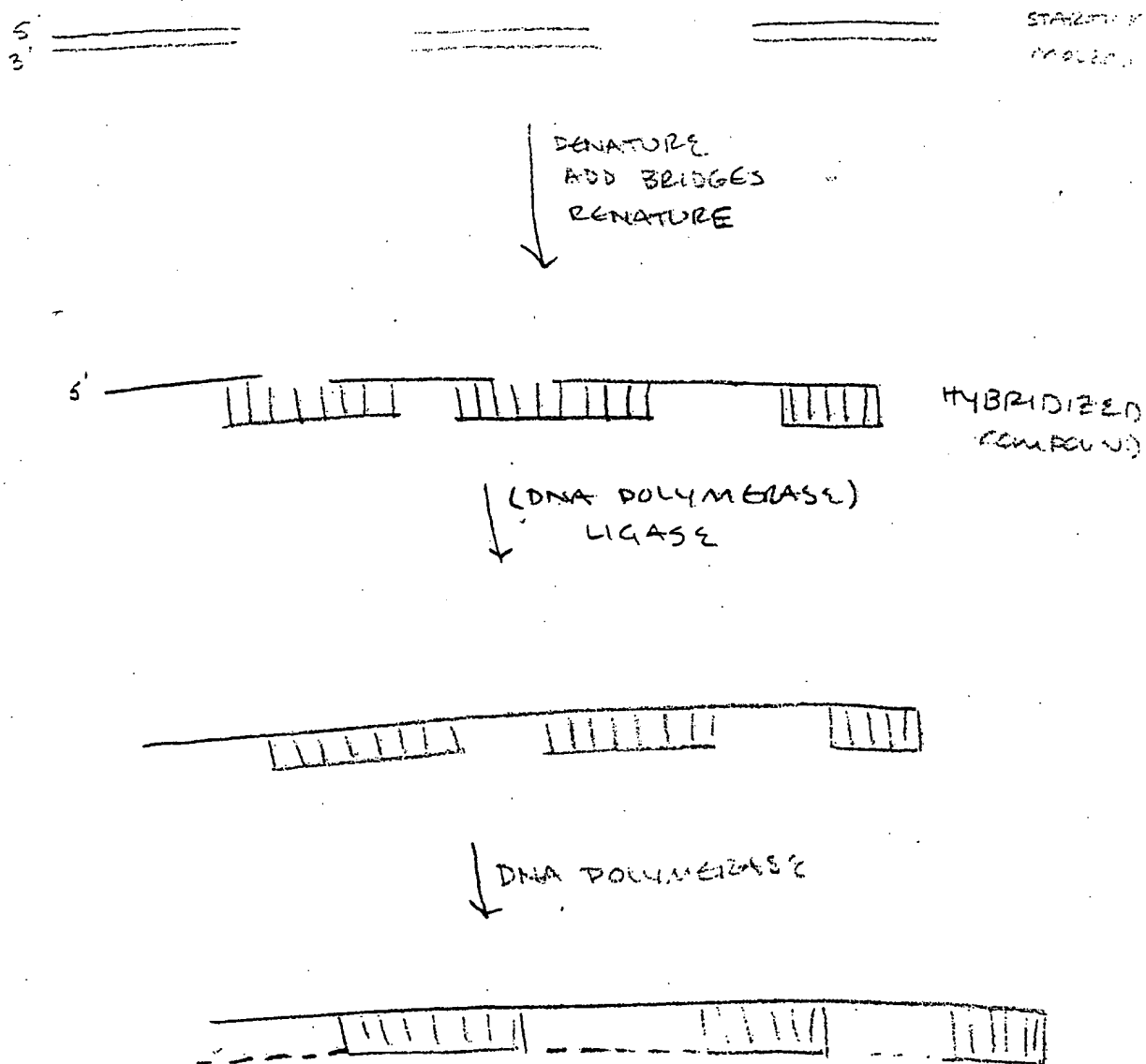
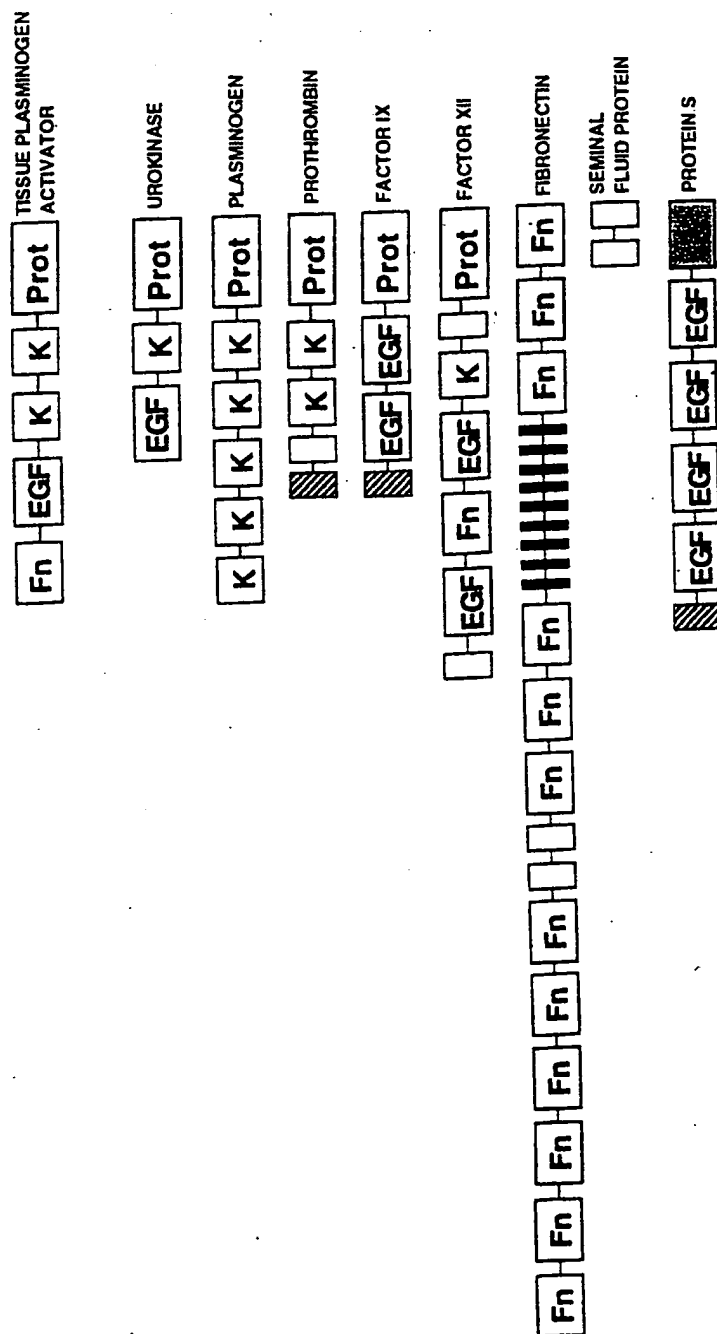


FIGURE 57

EXON SHUFFLING IN GENE EVOLUTION



10

Drugs Synthesized by Polyketide Synthases

Azithromycin	Idarubicin (Idamycin)
Clarithromycin	Amphotericin B
Erythromycin	Candididin
Dalfopristin	Griseofulvin
Josamycin	Nystatin/Mycostatin
Minocycline (Dynacil)	Spiramycin
Miokamycin	Mevacor (Lovastatin)
Mycinamicin	Mevastatin (Compactin)
Oleandomycin	Pravastatin
Pristinamycin	Zocor
Pseudomonic acid	Zearalenone
Rifamycins (Rifampin)	Ascomycin (Immunomycin)
Rokitamycin (Ricamycin)	FK506
Roxithromycin	Sirolimus (Rapamycin)
Tetracyclines	Avermectin
Aclarubicin (aclacinomycin)	Doramectin
Adriamycin (Doxorubicin)	Lasalocid A
Chromomycin	Milbemycin
Daunorubicin	Monensin
Enediynes	Tylosin

FIGURE 7

9/37

WO 98/01546

PCT/GB97/01819

2/40

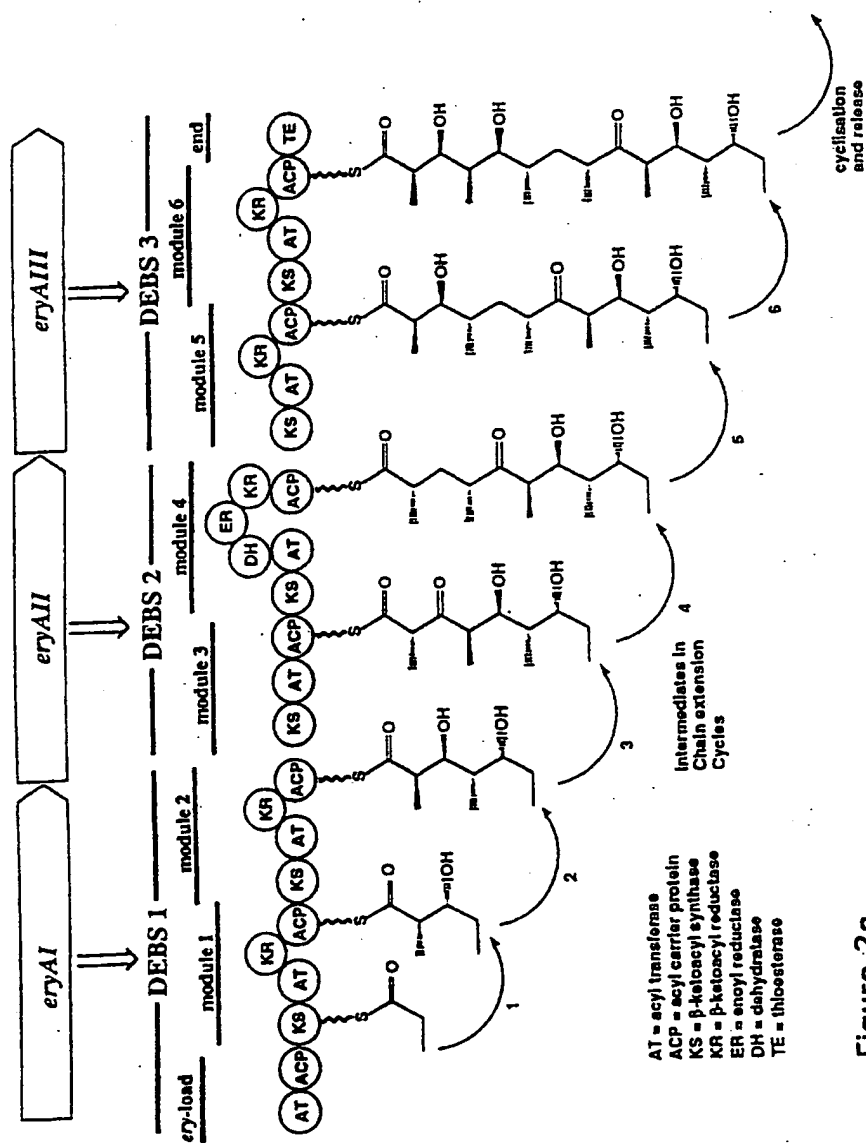


Figure 2a

10/37

4/40

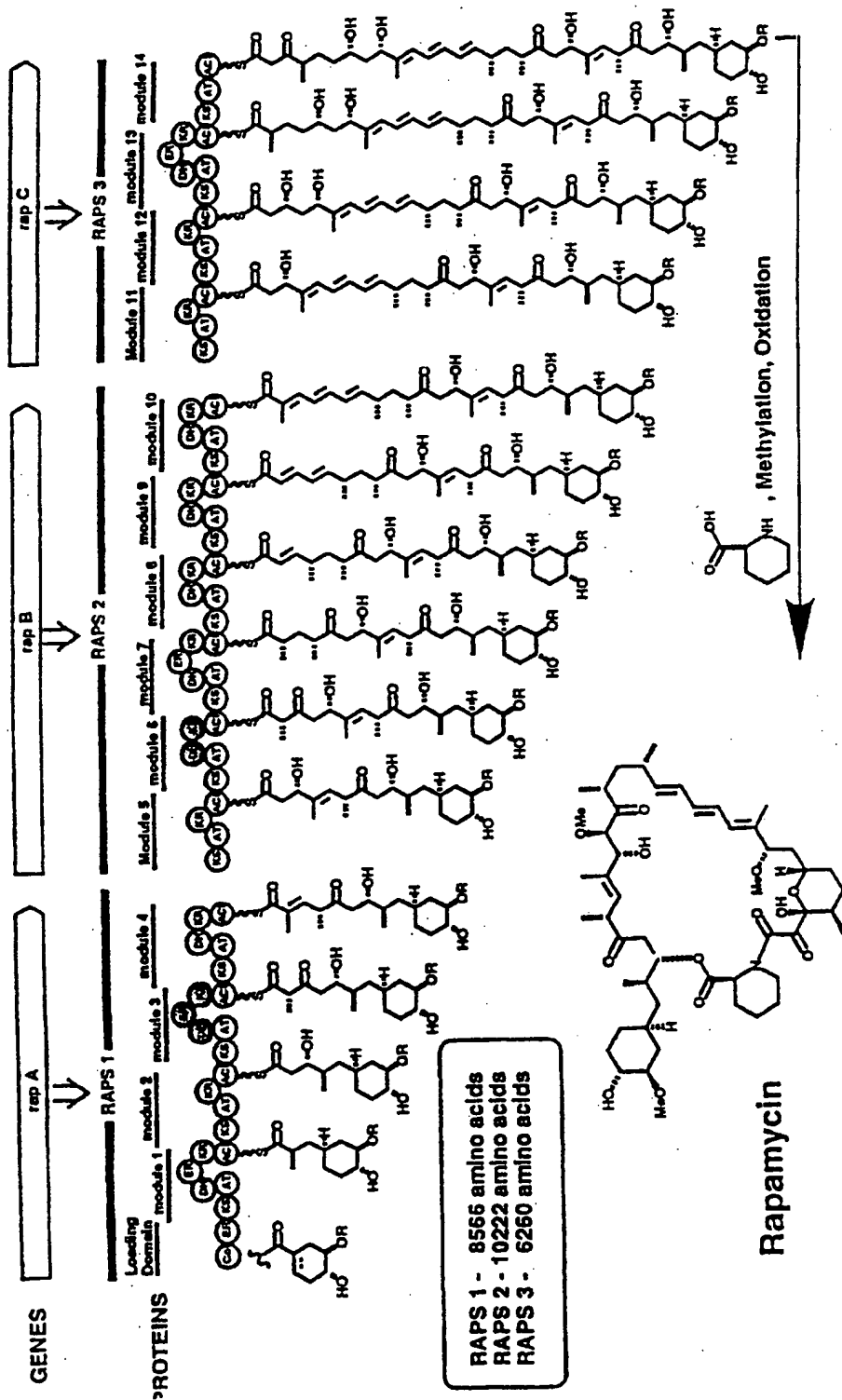


Figure 3

SUBSTITUTE SHEET (RULE 26)

FIGURE 9

11/37

MODULAR POLYKETIDE SYNTHASE GENES

<u>GENE</u>										<u>DRUG</u>									
DEBS										ERYTHROMYCIN									
AT	ACP	KS	AT	KR	ACP	KS	AT	KR	ACP										
RAPS										RAPAMYCIN									
Lig	ER	ACP	KS	AT	DH	ER	KR	ACP	KS										

POLYKETIDES

41 APPROVED DRUGS

8 THERAPEUTIC AREAS INCLUDING:

ANTIBACTERIAL, ANTICANCER, ANTIFUNGAL, ANTIRICKETTSIAL,
CHOLESTEROL-LOWERING, IMMUNOSUPPRESSANT

FIGURE 10

12/37

Drugs Synthesized by Peptide Synthetases

Penicillin
Cephalosporins
Clavulanic acid
Blalaphos
Pristinamycins
Actinomycin
Viridogrisein
Enniatins
A47934
Vancomycin
Teichoplanin
Ardacin
Surfactin
Bacitracin
Cyclosporin

FIGURE 11.

MODULAR PEPTIDE SYNTHETASE GENES

GENE DRUG

ACVS



PENICILLIN

Aad Cys D-Val

CY



CYCLOSPORIN

D-Ala N-Me-Leu N-Me-Leu

PEPTIDES

14 APPROVED DRUGS

4 THERAPEUTIC AREAS INCLUDING:

ANTIBACTERIAL, ANTICANCER, ANTIVIRAL, IMMUNOSUPPRESSANT

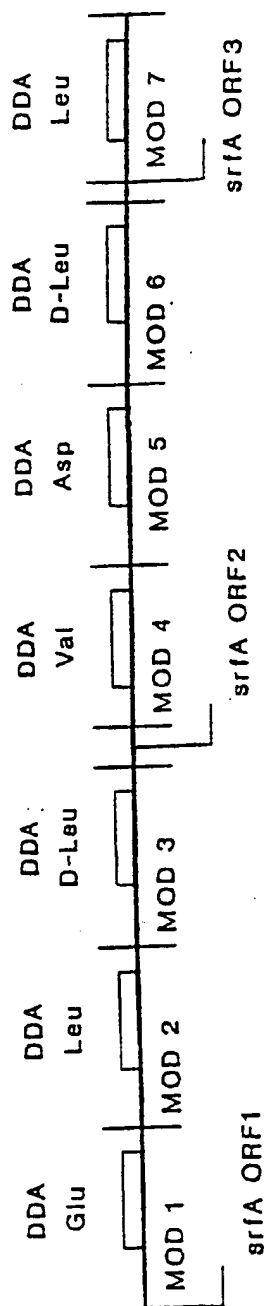
U.S. Patent

Jul. 29, 1997

Sheet 1 of 8

5,652,116

Fig. 1

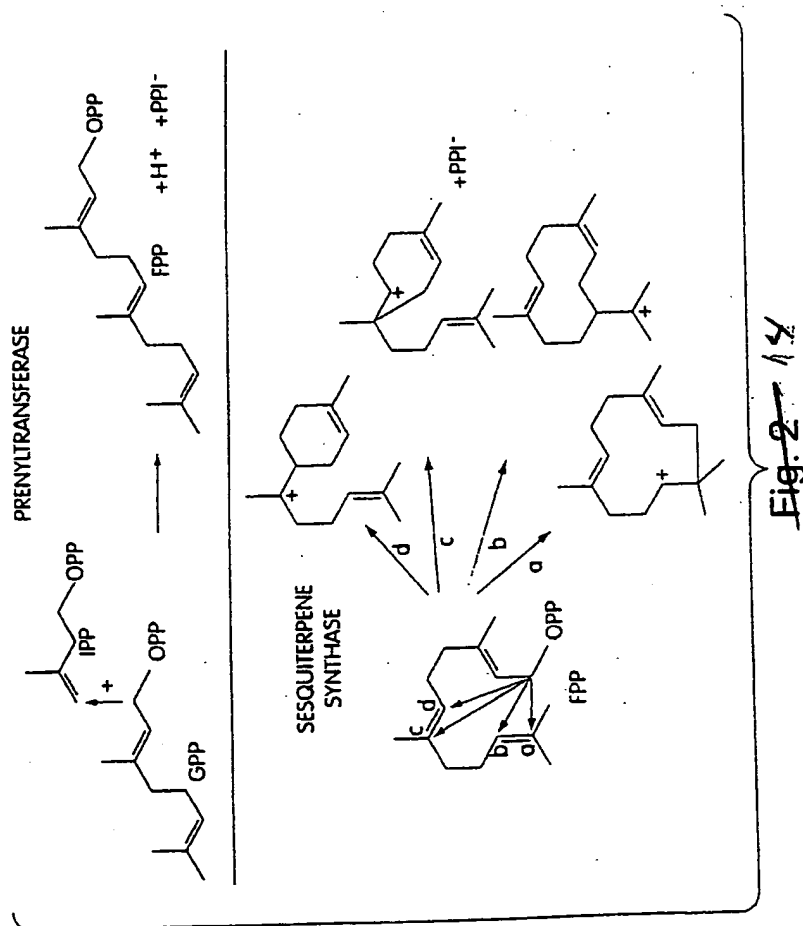


U.S. Patent

Oct. 20, 1998

Sheet 2 of 9

5,824,774

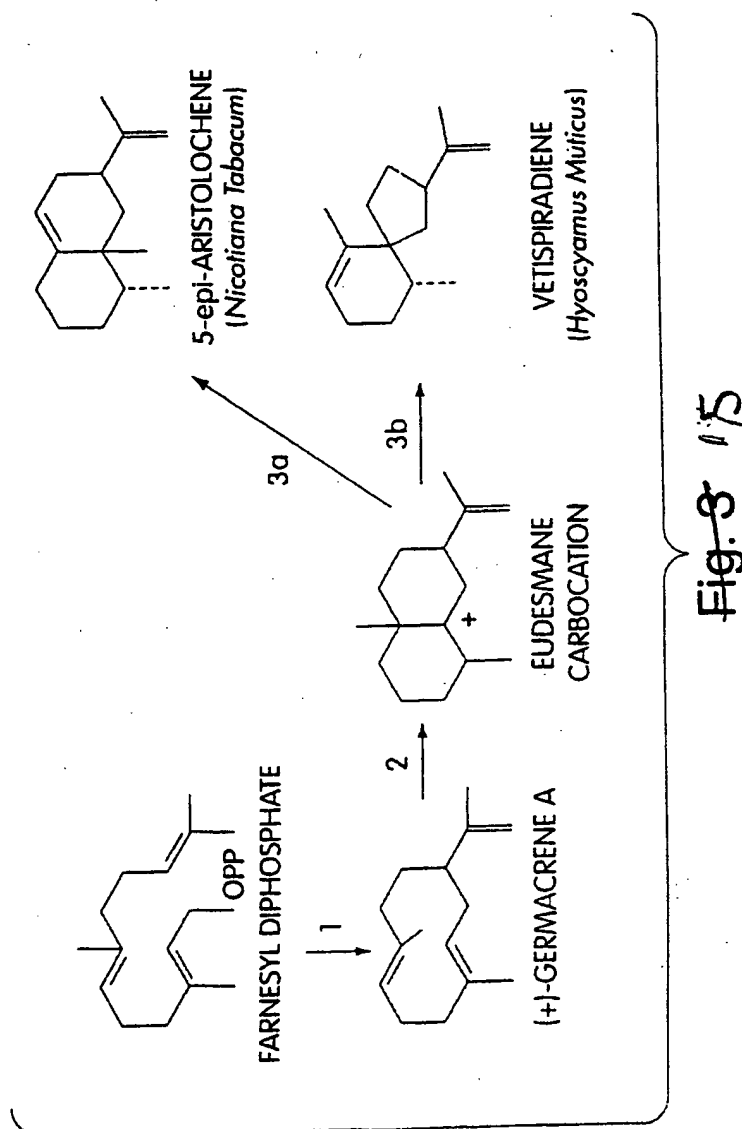


U.S. Patent

Oct. 20, 1998

Sheet 3 of 9

5,824,774

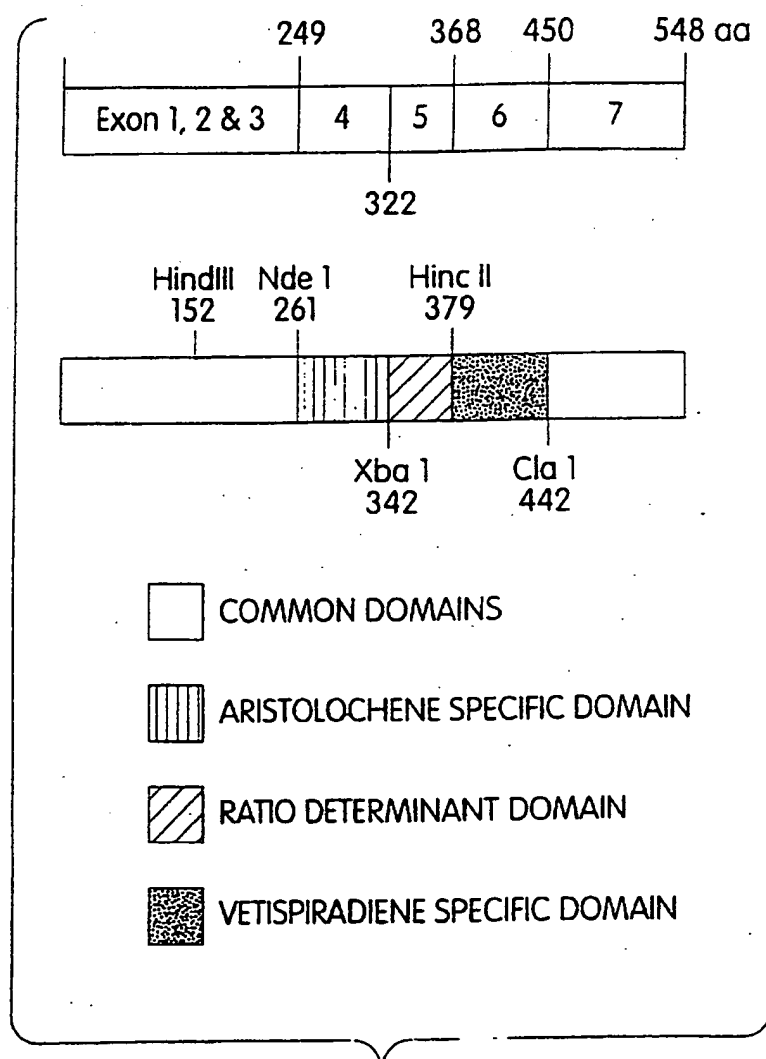


U.S. Patent

Oct. 20, 1998

Sheet 6 of 9

5,824,774

~~Fig. 5~~ 16

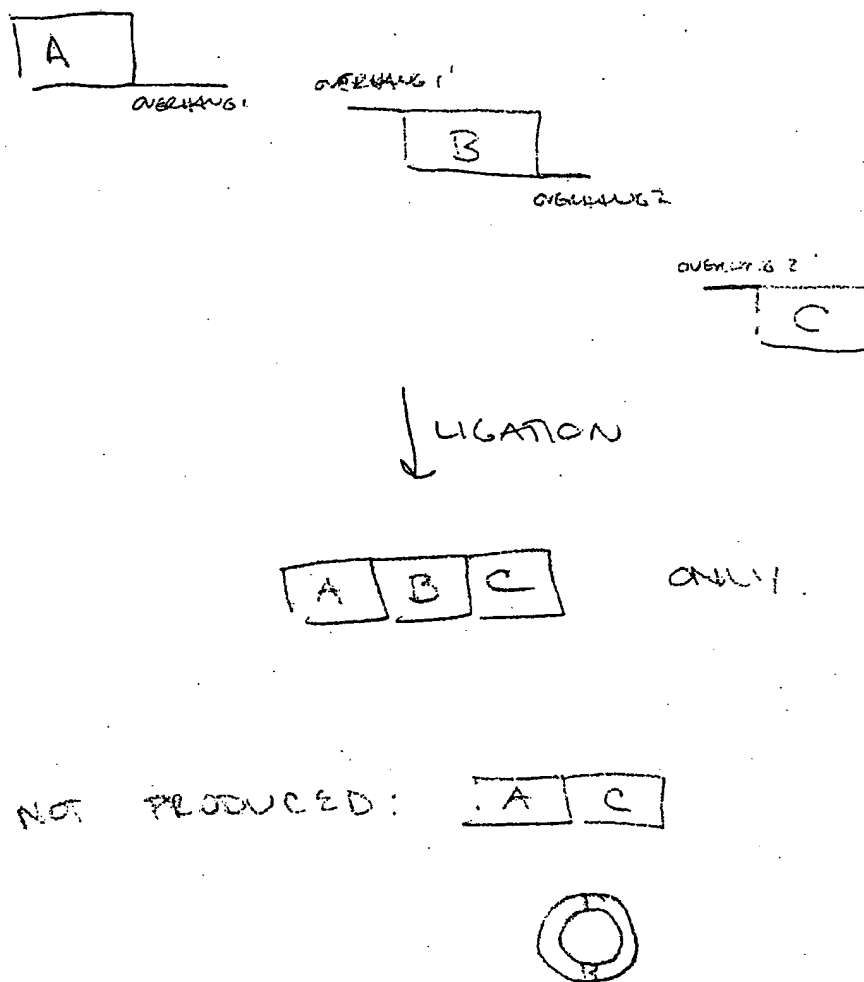


FIGURE 17

19 / 37

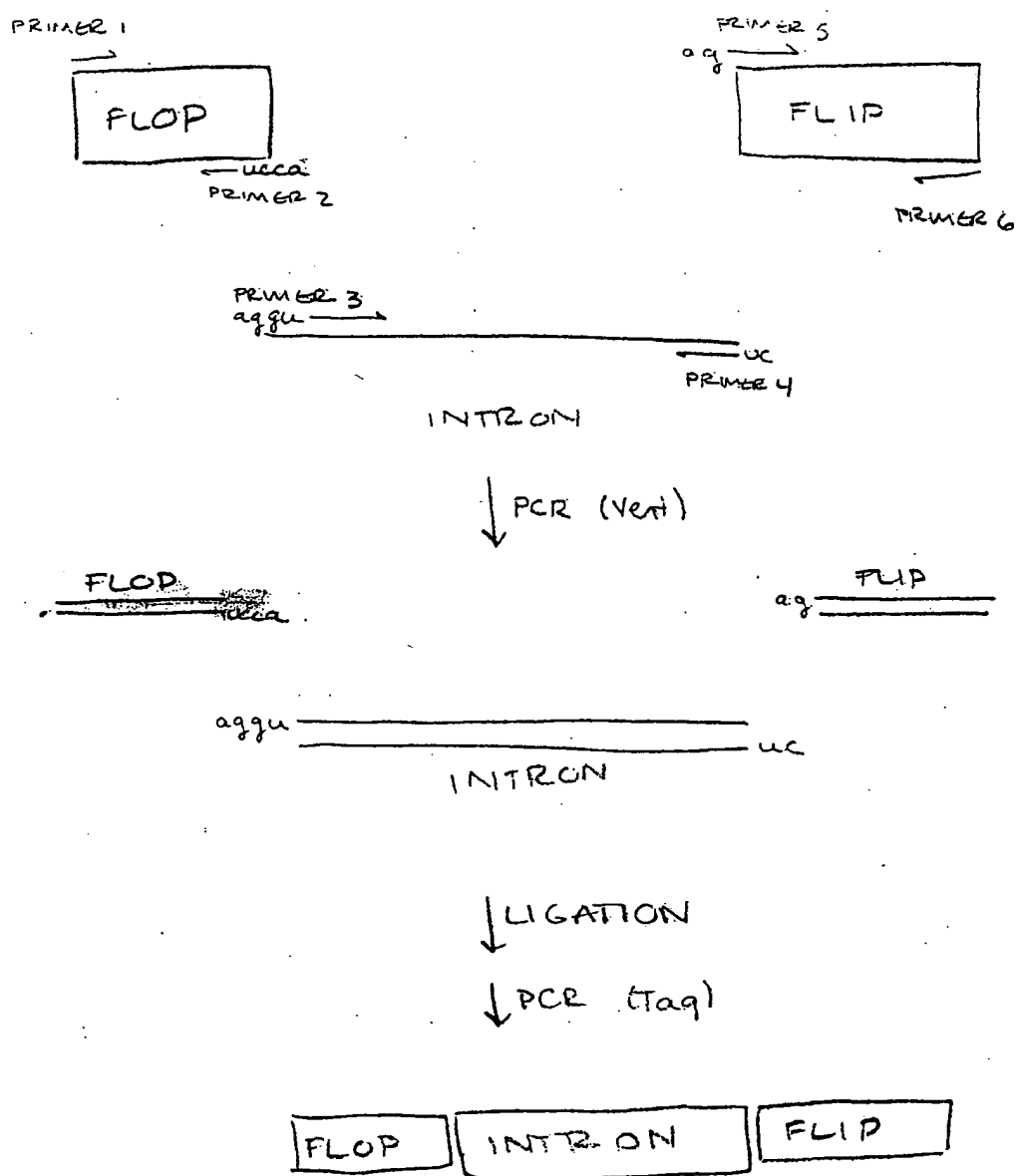


FIGURE 18

20 / 37

NCBI Entrez Nucleotide QUERY **BLAST Entrez**

Other Formats: **FASTA** **Graphic**

Links: **Related Sequences**

LOCUS HSGRSFLIP 188 bp RNA PRI 16-MAR-1992
DEFINITION H.sapiens mRNA for GLU-R2 glutamate receptor subunit, 'flip' exon.
ACCESSION X64829
NID g31910
KEYWORDS glutamate receptor subunit.
SOURCE human.
ORGANISM Homo sapiens
Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
Vertebrata; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 188)
AUTHORS Cheethan,M.E.
TITLE Direct Submission
JOURNAL Submitted (05-MAR-1992) M.E. Cheethan, Institute of Psychiatry, De
Crespigny park, Denmark Hill, London SE5 8AF, UK
REFERENCE 2 (bases 1 to 188)
AUTHORS McLaughlin,D.P., Cheethan,M.E. and Kerwin,R.W.
TITLE Nucleotide sequence of Flip and Flop exons of a human glutamate
receptor gene
JOURNAL Unpublished
FEATURES Location/Qualifiers
source 1..188
/organism="Homo sapiens"
/db_xref="taxon:9606"
/haplotype="diploid"
/tissue_type="brain (frontal cortex)"
exon 10..123
/note="alternatively spliced 'flip' exon"
misc_feature 153..188
/note="TMIV region"
BASE COUNT 58 a 37 c 50 g 43 t
ORIGIN
1 tcattaggaa cccagtaaa tcttgagta ttgaaactca gtgagcaagg cgtcttagac
61 aagctgaaaa acaaatggtg gtacgataaa ggtgaatgtg gagccaagga ctctggaagt
121 aagaaaagac cagtgccttc agtctgagca acgttgctgg agtattctac atccttgctg
181 ggggcctt
//

Save the above report in **Macintosh** **Text** format

FIGURE 119A

21 / 37

Entrez		Nucleotide QUERY		BLAST Entrez	
Other Formats:		FASTA	Graphic		
Links:		Related Sequences			
LOCUS	HSGRSFLOP 190 bp RNA PRI 16-MAR-1992				
DEFINITION	H.sapiens mRNA for GLU-R2 glutamate receptor subunit, 'flop' exon.				
ACCESSION	X64830				
NID	g31911				
KEYWORDS	glutamate receptor subunit.				
SOURCE	human.				
ORGANISM	Homo sapiens				
	Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata; Vertebrata; Eutheria; Primates; Catarrhini; Hominidae; Homo.				
REFERENCE	1 (bases 1 to 190)				
AUTHORS	Cheethman, M.E.				
TITLE	Direct Submission				
JOURNAL	Submitted (05-MAR-1992) M.E. Cheethman, Institute of Psychiatry, De Crespigny park, Denmark Hill, London SE5 8AF, UK				
REFERENCE	2 (bases 1 to 190)				
AUTHORS	McLaughlin, D.P., Cheethan, M.E. and Kerwin, R.W.				
TITLE	Nucleotide sequence of Flip and Flop exons of a human glutamate receptor gene				
JOURNAL	Unpublished				
FEATURES	Location/Qualifiers				
source	1..190				
	/organism="Homo sapiens"				
	/db_xref="taxon:9606"				
	/haplotype="diploid"				
	/tissue_type="brain (frontal cortex)"				
exon	10..123				
	/note="alternatively spliced 'flop' exon"				
misc_feature	155..190				
	/note="TMIV region"				
BASE COUNT	57 a 37 c 56 g 40 t				
ORIGIN	1 tcattaggaa atgcggttaa cctgcagta ctaaaactga atgaacaagg cctgttgga				
	61 aaattgaaaa acaaatggtg gtacgacaaa ggagagtgcg gcagcggggg aggtgattcc				
	121 aagggaaaag accagtgcgc tcagtctgag caacgttgct ggagtattct acatccttgc				
	181 cgggggcctt				
//					

Save the above report in format

FIGURE 19B...

22 / 37

Individual PCR amplified fragments and
the PCR amplified chimeric construct

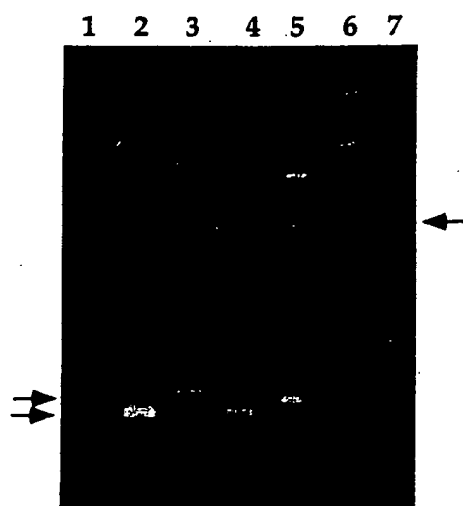


FIGURE 24.

The Flop/ β -globin intron/ Flip chimera ligation
site sequences

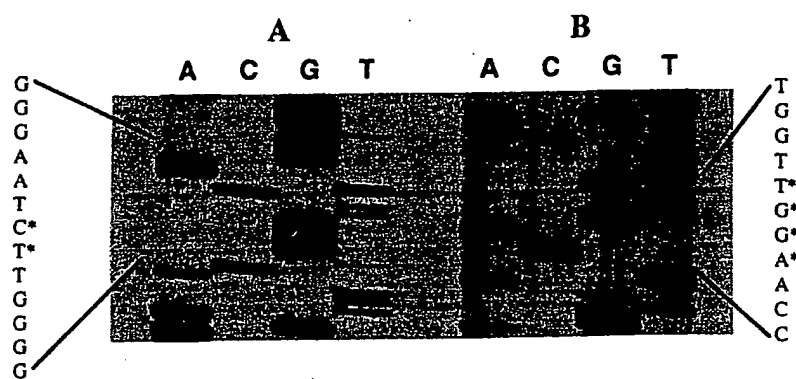


FIGURE 2d

24 / 37

DNA Strider 1.0 ### Tuesday, December 1, 1998 5:01:31 PM

FBF* sequence -> Full Restriction Map

DNA sequence 360 b.p. AAATGCGGTAA ... TCGGAAGTAAG linear

Positions of Restriction Endonucleases sites (unique sites underlined)

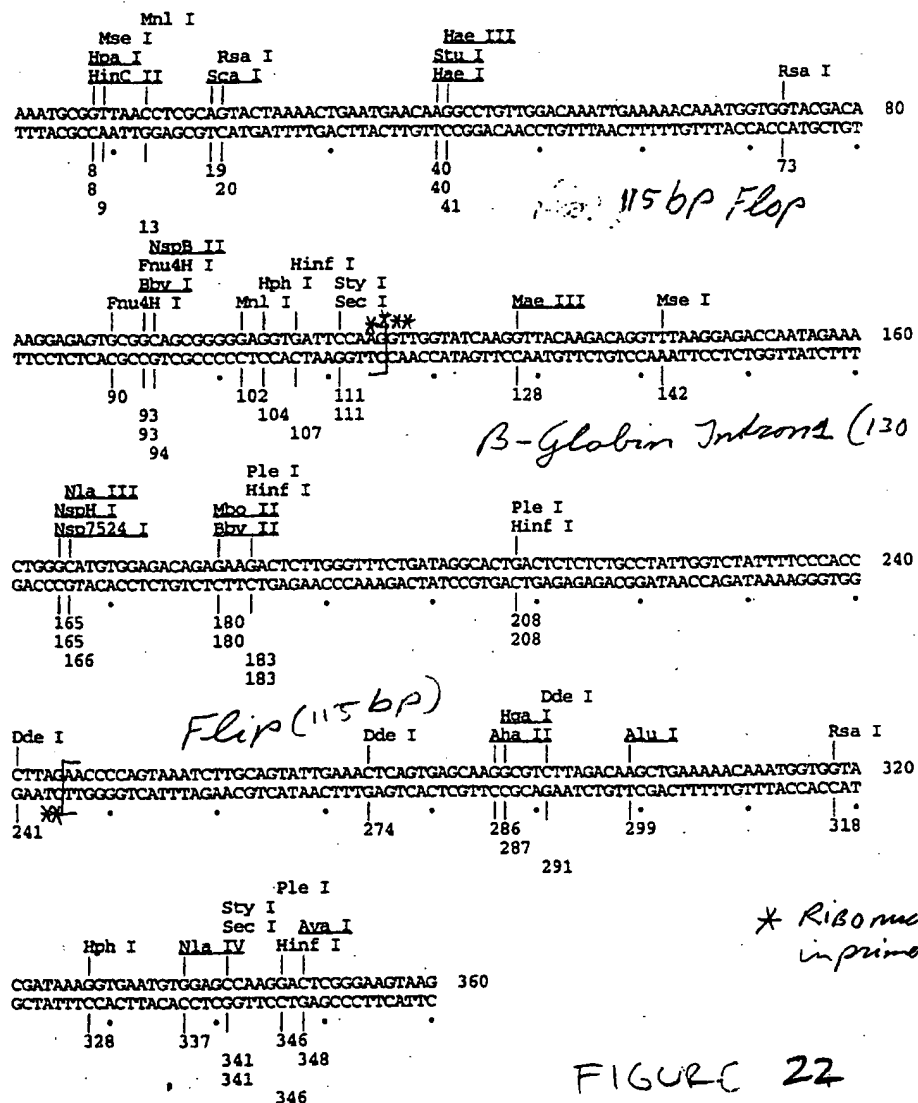


FIGURE 22

Restriction Endonucleases site usage

Aat II	-	BstE II	-	Hinf I	4	PpuM I	-
Acc I	-	BstN I	-	HinP I	-	Pst I	-
Afl I	-	BstU I	-	Hpa I	1	Pvu I	-
Afl III	-	BstX I	-	Hpa II	-	Pvu II	-
Aha II	1	BstY I	-	Hph I	2	Rsa I	3
Alu I	1	Bsu36 I	-	Kpn I	-	Rsr II	-
Alw I	-	Cfr10 I	-	Mae I	-	Sac I	-
AlwN I	-	Cla I	-	Mae II	-	Sac II	-

Apa I	-	Dde I	-	Mae III	1	Sal I	-
ApaL I	-	Dpn I	-	Mbo I	-	Sau3A I	-
Ase I	-	Dra I	-	Mbo II	1	Sau96 I	-
Asp718	-	Dra III	-	Mlu I	-	Sca I	1
Ava I	1	Eae I	-	Mnl I	2	ScrF I	-
Ava II	-	Eag I	-	Mse I	2	Sec I	2
Avr II	-	Eco47 III	-	Msp I	-	SfaN I	-
Bal I	-	EcoN I	-	Nae I	-	Sfi I	-
BamH I	-	EcoO109 I	-	Nar I	-	Sma I	-
Ban I	-	EcoR I	-	Nci I	-	SnaB I	-
Ban II	-	EcoR II	-	Nco I	-	Spe I	-
Ebe I	-	EcoR V	-	Nde I	-	Sph I	-
Ebv I	1	Esp I	-	Nhe I	-	Spl I	-
Ebv II	1	Fnu4H I	2	Nla III	1	Ssp I	-
Ecl I	-	Fok I	-	Nla IV	1	Stu I	1
Ecn I	-	Fsp I	-	Not I	-	Sty I	2
Egl I	-	Gdi II	-	Nru I	-	Taq I	-
Egl II	-	Hae I	1	Nsi I	-	TchIII I	-
Esm I	-	Hae II	-	Nsp7524 I	1	TchIII II	-
Esp1286 I	-	Hae III	1	NspB II	1	Xba I	-
EspH I	-	Hga I	1	NspH I	1	Xca I	-
Espm I	-	HgiA I	-	PaeR7 I	-	Xho I	-
Espm II	-	Hha I	-	PflM I	-	Xma I	-
Essh II	-	Hinc II	1	Ple I	3	Xmn I	-
EstB I	-	Hind III	-				

Enzyme	Site	Use	Site position	(Fragment length)	Fragment order	
Aha II	gr/cgyc	1	1(285) 1	286(75) 2		
Alu I	ag/ct	1	1(298) 1	299(62) 2		
Ava I	c/ycgrrg	1	1(347) 1	348(13) 2		
Bbv I	gcagc	8/12	1(92) 2	93(268) 1		
Bbv II	gaagac	2/6	1(179) 2	180(181) 1		
Hae I	wgg/ccw	1	1(39) 2	40(321) 1		
Hae III	gg/cc	1	1(40) 2	41(320) 1		
Hga I	gacgc	5/10	1(286) 1	287(74) 2		
HinC II	gty/rac	1	1(7) 2	8(353) 1		
Hpa I	gtt/aac	1	1(7) 2	8(353) 1		
Mae III	/gtnac	1	1(127) 2	128(233) 1		
Mbo II	gaaga	8/7	1(179) 2	180(181) 1		
Nla III	catg/	1	1(165) 2	166(195) 1		
Nla IV	ggn/ncc	1	1(336) 1	337(24) 2		
Nsp7524 I	r/catgy	1	1(164) 2	165(196) 1		
NspB II	cmg/ckg	1	1(93) 2	94(267) 1		
NspH I	rcatg/y	1	1(164) 2	165(196) 1		
Sca I	agt/act	1	1(18) 2	19(342) 1		
Stu I	agg/cct	1	1(39) 2	40(321) 1		
Pvu4H I	gc/ngc	2	1(89) 2	90(3) 3	93(268) 1	
Hph I	ggtga	8/7	1(103) 2	104(224) 1	328(33) 3	
Hnl I	cctc	7/7	1(12) 3	13(89) 2	102(259) 1	
Mse I	t/taa	2	1(8) 3	9(133) 2	142(219) 1	
Sec I	c/cnngg	2	1(110) 2	111(230) 1	341(20) 3	
Sty I	c/cwggg	2	1(110) 2	111(230) 1	341(20) 3	
Dde I	c/tnag	3	1(240) 1	241(33) 3	274(17) 4	291(70) 2
Ple I	gagtc	4/5	1(182) 1	183(25) 3	208(138) 2	346(15) 4
Rsa I	gt/ac	3	1(19) 4	20(53) 2	73(245) 1	318(43) 3
Hinf I	g/antc	4	1(106) 2	107(76) 3	183(25) 4	208(138) 1
			346(15) 5			

44 sites found

No Sites found for the following Restriction Endonucleases

Aat II	gaagt/c	Cfr10 I	r/ccggy	Nha I	g/ctagc
Acc I	gt/mkac	Cla I	at/cgat	Not I	gc/ggccgc
Afl II	c/ttaag	Dpn I	ga/tc	Nru I	tc/gcga
Afl III	a/crygt	Dra I	t/tc/aaa	Nsi I	atgca/t
Alw I	ggatc	Dra III	caacnnn/gtg	Paek7 I	c/tcgag
AlwN I	caacnnn/ctg	Rae I	y/ggccr	PflM I	caacnnnn/ntgg
Apa I	gggcc/c	Eag I	c/ggccg	PpuM I	rg/gwccy
Apal I	g/tgcac	Eco47 III	agc/gct	Pst I	ctgca/g
Ase I	at/taat	EcoN I	cctnn/nnnaggy	Pvu I	cgat/cg

26 / 37

Asp718	g/gtacc	EcoO11 I	rg/gnccy	Pvu II	cag/ctg
Ava II	g/gwcc	EcoR I	g/aattc	Rsr II	cg/gwccg
Avr II	c/ctagg	EcoR II	/ccwgg	Sac I	gagct/c
Bal I	tgg/cca	EcoR V	gat/atc	Sac II	ccgc/gg
BamH I	g/gatcc	Esp I	gc/tnagc	Sal I	g/tcgac
Ban I	g/gyrcc	Fok I	ggatg	Sau3A I	/gatc
Ban II	grgcy/c	Fsp I	tgc/gca	Sau96 I	g/gncc
Bbe I	ggcgc/c	Gdi II	yggccg	ScrF I	cc/ngg
Bcl I	t/gatca	Hae II	rgcgc/y	SfaN I	gcac 5/9
Bcn I	ccs/gg	HgiA I	gwgwc/c	Sfi I	ggccnnnn/nggcc
Bgl I	gccnnnn/nggc	Hha I	gcg/c	Sma I	ccc/ggg
Bgl II	a/gatct	HinD III	a/agctt	SnaB I	tac/gta
Bsm I	gaatgc 1/-1	HinP I	g/cgc	Spe I	a/ctagt
Bsp1286 I	gdgch/c	Hpa II	c/cgg	Sph I	gcatg/c
BspH I	t/catga	Kpn I	ggtac/c	Spl I	c/gtacg
BspM I	acctgc 4/8	Mae I	c/tag	Ssp I	aat/att
BspM II	t/ccgga	Mae II	a/cgt	Taq I	t/cga
BssH II	g/cgcgc	Mbo I	/gatc	Tth111 I	gacn/nngtc
BstB I	tt/cgaa	Mlu I	a/cgcgt	Tth111 II	caarca 11/9
BstE II	g/gtnacc	Msp I	c/cgg	Xba I	t/ctaga
BstN I	cc/wgg	Nae I	gcc/ggc	Xca I	gta/tac
BstU I	cg/cg	Nar I	gg/cgcc	Xho I	c/tcgag
BstX I	ccannnnn/ntgg	Nci I	cc/sgg	Xma I	c/ccggg
BstY I	r/gatcy	Nco I	c/catgg	Xmn I	gaann/nnttc
Bsu36 I	cc/tnagg	Nde I	ca/tatg		

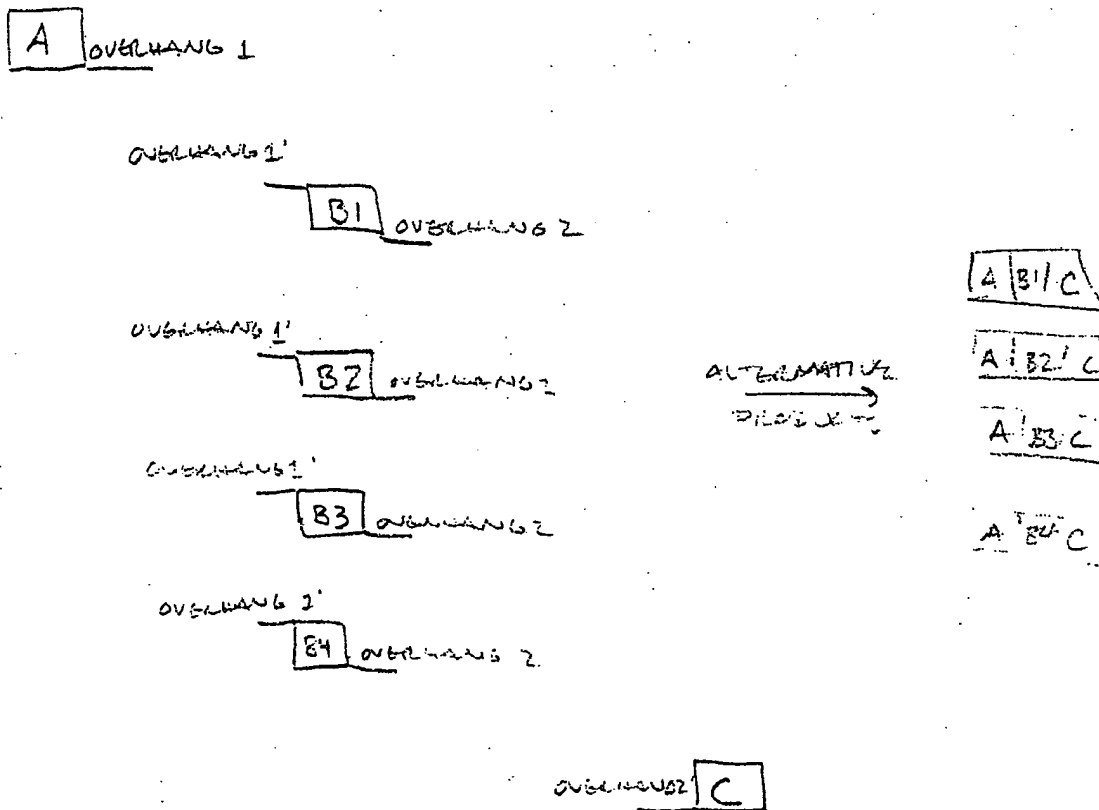


FIGURE 23

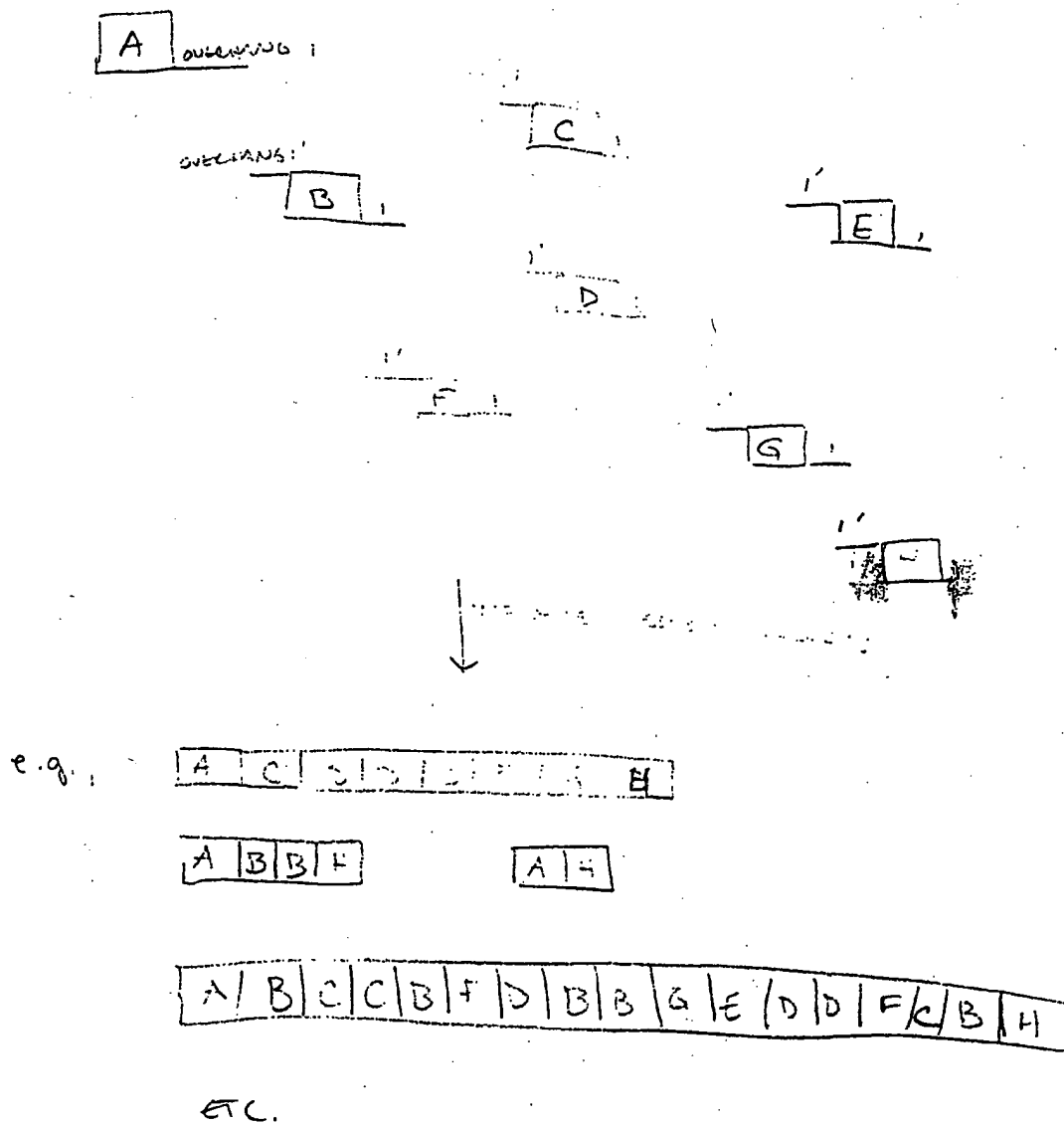
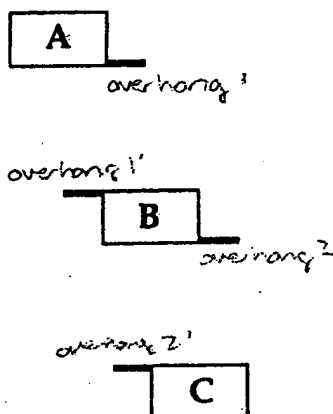


FIGURE 24

LIBRARY ASSEMBLY USING RNA/DNA CHIMERIC OLIGOS



COMBINATORIAL POTENTIAL

$$10 \times 3 = 30$$

$$10^3 = 1000$$

FIGURE 25

LIBRARY ASSEMBLY USING RNA/DNA CHIMERIC OLIGOS AND INTRON SPLICING

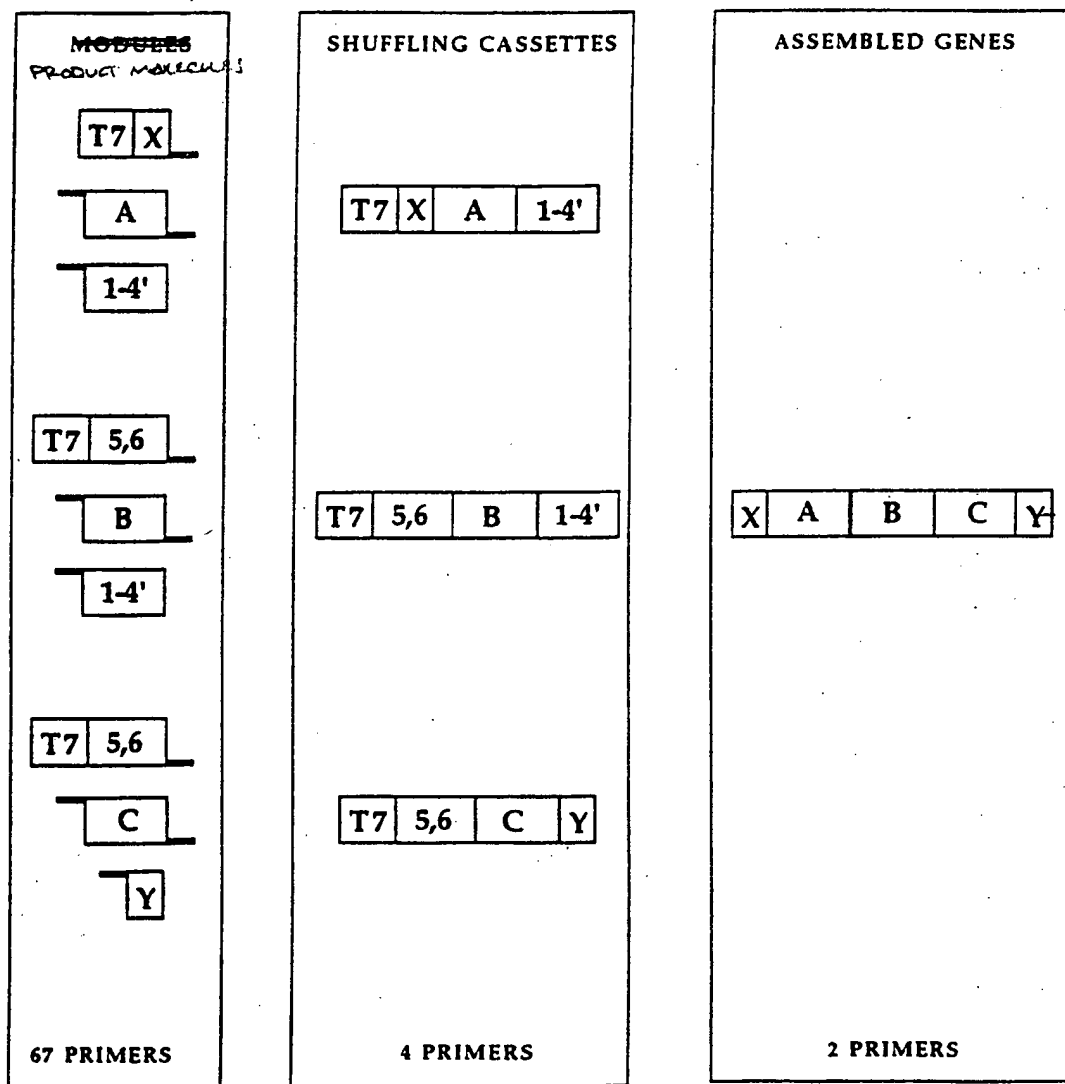


FIGURE 26

31 / 37

AUTOMATED GENE ASSEMBLY, SCREENING AND INFORMATICS

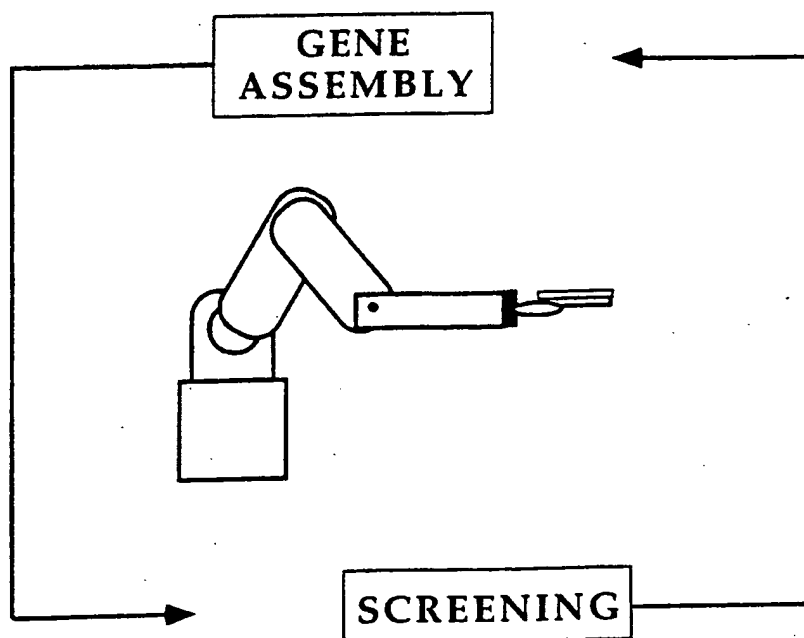


FIGURE 27

32 / 37

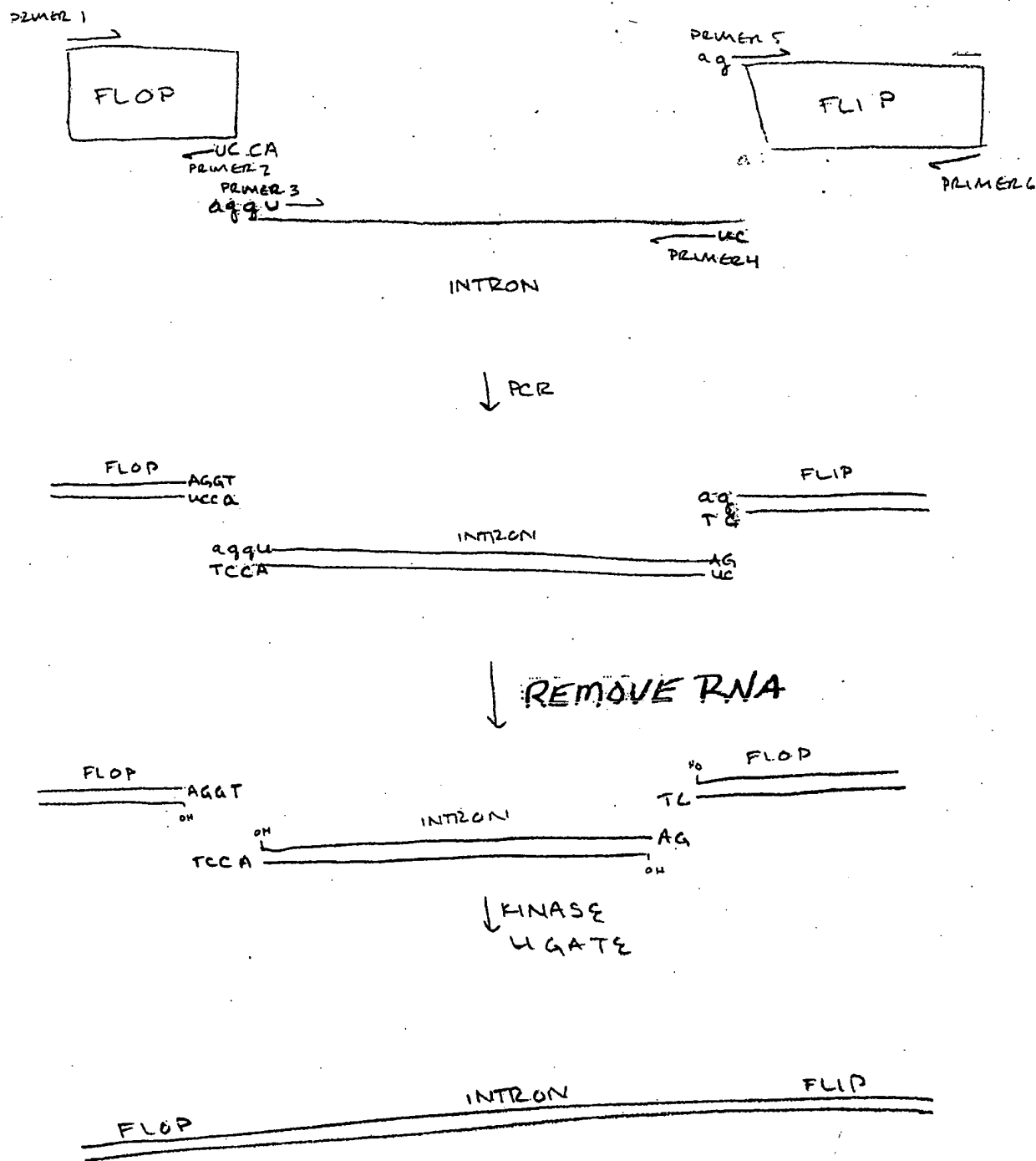


FIGURE 28

33 / 37

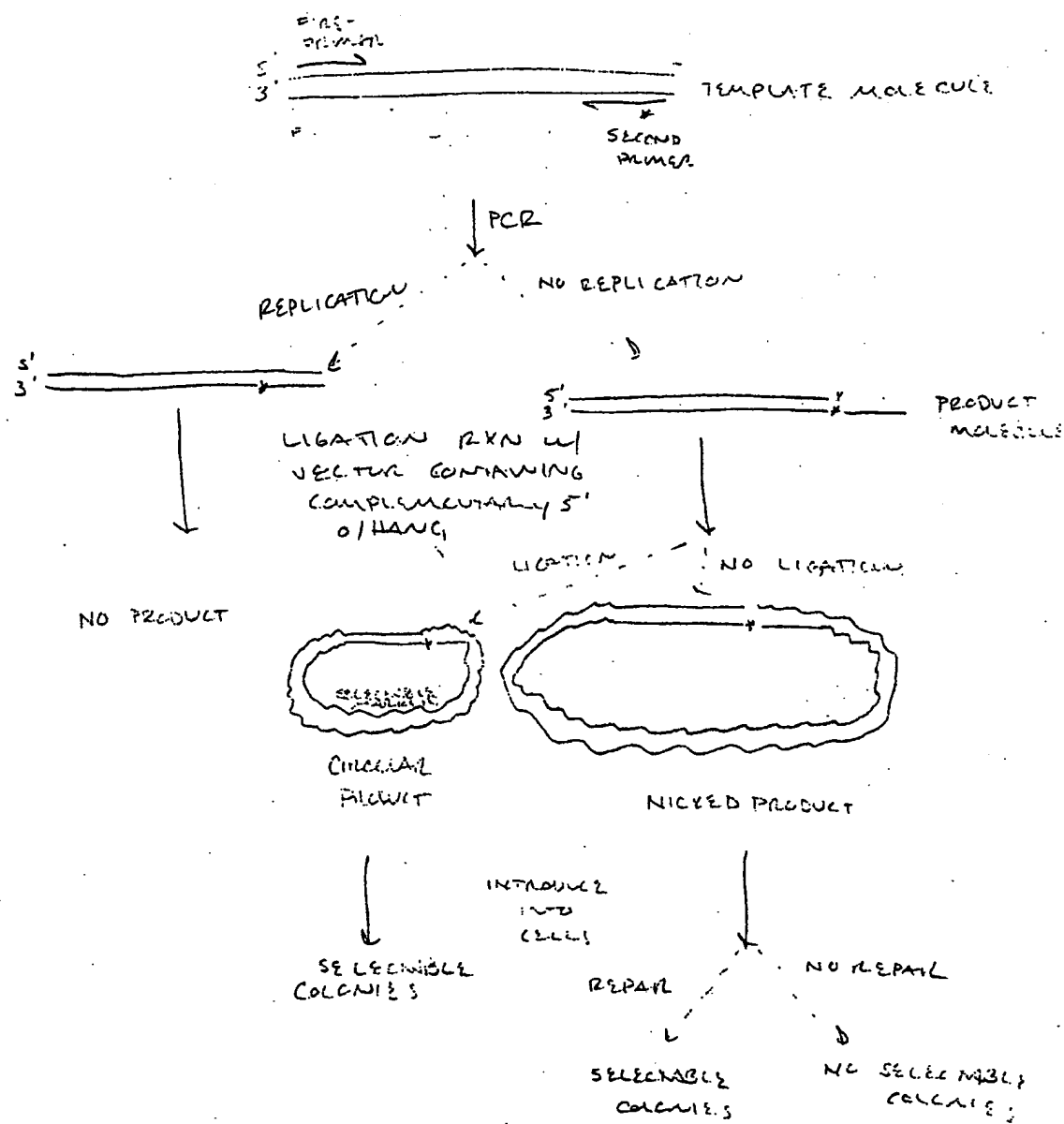


FIGURE 29

EXON SHUFFLING WITH HETEROLOGOUS INTRONS

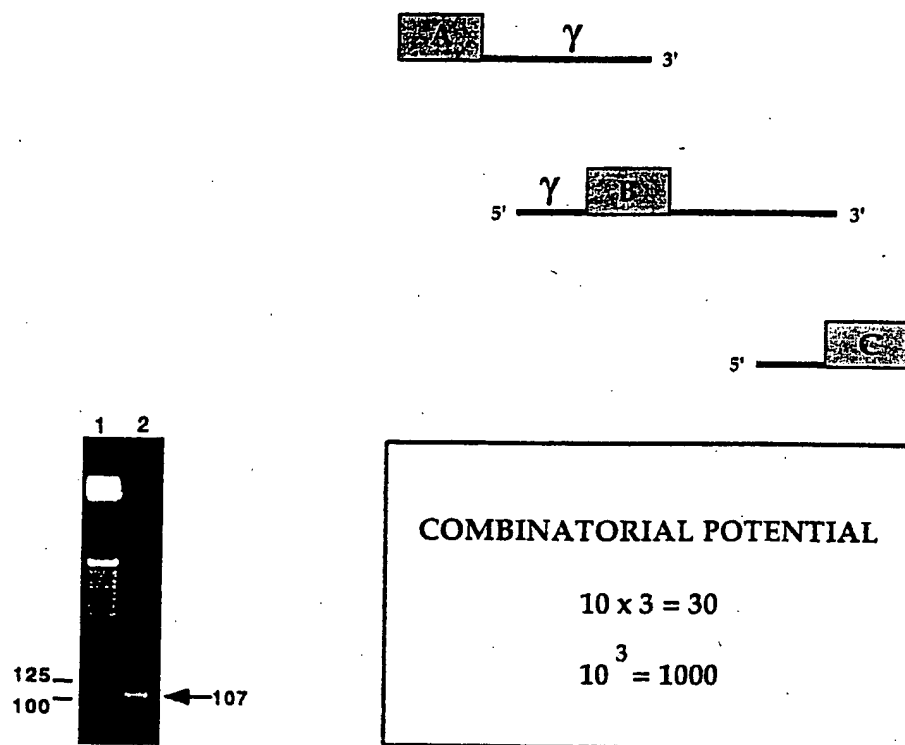


FIGURE 30

SHUFFLING VECTORS

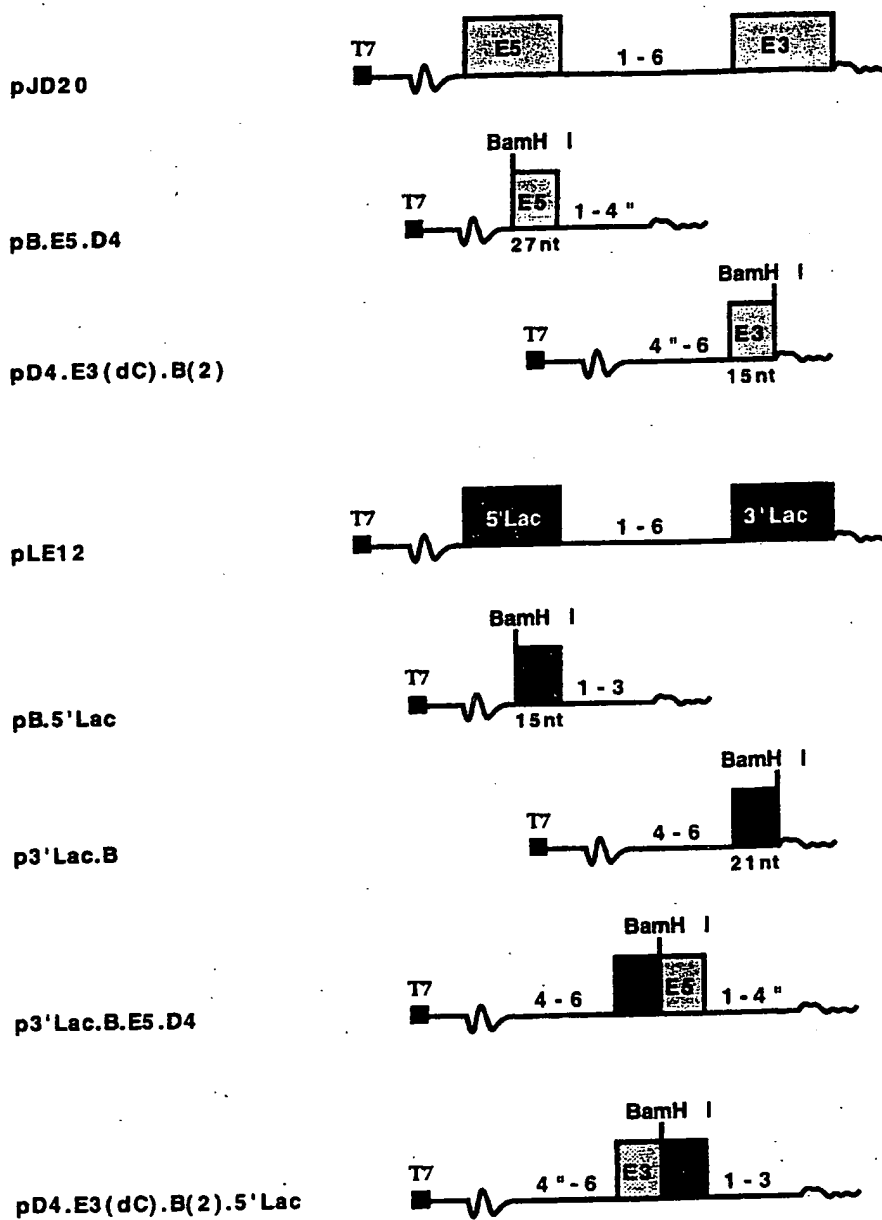


FIGURE 3

Fig. 32



Fig. 33



PARENTAL GENES



CHIMERIC GENES



FIGURE 34